

Problemas resueltos de estadística aplicada a las ciencias sociales

Pablo Juan Verdoy
Modesto Joaquín Beltrán
María José Peris

Problemas resueltos de estadística aplicada a las ciencias sociales

Pablo Juan Verdoy
Modesto Joaquín Beltrán
María José Peris



DEPARTAMENT DE MATEMÀTIQUES

■ Codis d'assignatura RA10, RL0906

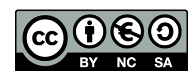
Edita: Publicacions de la Universitat Jaume I. Servei de Comunicació i Publicacions
Campus del Riu Sec. Edifici Rectorat i Serveis Centrals. 12071 Castelló de la Plana
http://www.tenda.uji.es e-mail: publicacions@uji.es

Col·lecció Sapientia, 100
www.sapientia.uji.es
Primera edició, 2015

ISBN: 978-84-15444-38-1



Publicacions de la Universitat Jaume I és una editorial membre de l'UNE, cosa que en garanteix la difusió de les obres en els àmbits nacional i internacional. www.une.es



ÍNDICE

P. Juan Verdoy / M. J. Beltrán / M. J. Peris - ISBN: 978-84-15444-38-1 Problemas resueltos de estadística aplicada a las ciencias sociales - UJI - DOI: http://dx.doi.org/10.6035/Sapientia100

Prólogo

La estadística es una ciencia con base matemática referente a la recogida, análisis e interpretación de datos que busca explicar condiciones regulares en fenómenos de tipo aleatorio. Es transversal a una amplia variedad de disciplinas, desde la física hasta las ciencias sociales, desde las ciencias de la salud hasta el control de calidad, y es usada para la toma de decisiones en áreas de negocios e instituciones gubernamentales.

Podemos considerar dos ramas en la Estadística:

a) La estadística descriptiva, que se dedica a los métodos de recogida, descripción, visualización y resumen de datos originados a partir de los fenómenos en estudio. Los datos pueden ser resumidos numérica o gráficamente. Ejemplos básicos de parámetros estadísticos son: la media y la desviación estándar. Algunos ejemplos gráficos son: histograma, pirámide poblacional, clústeres, etc.

b) La inferencia estadística se dedica a la generación de los modelos, inferencias y predicciones asociadas a los fenómenos en cuestión teniendo en cuenta la aleatoriedad de las observaciones. Se usa para modelar patrones en los datos y extraer inferencias sobre la población de estudio. Estas inferencias pueden tomar la forma de respuestas a preguntas sí/no (prueba de hipótesis), estimaciones de características numéricas (estimación), pronósticos de futuras observaciones, descripciones de asociación (correlación) o modelización de relaciones entre variables (análisis de regresión). Otras técnicas de modelización incluyen ANOVA, series de tiempo y minería de datos.

Ambas ramas (descriptiva e inferencial) comprenden la estadística aplicada. Hay también una disciplina llamada estadística matemática, la cual hace referencia a las bases teóricas de la materia. La palabra «estadística» también se refiere al resultado de aplicar un algoritmo estadístico a un conjunto de datos, como en estadísticas económicas, estadísticas criminales, etc.

En su origen, por lo tanto, la estadística estuvo asociada a datos para ser utilizados por el gobierno y cuerpos administrativos (a menudo centralizados). La colección de datos sobre estados y localidades continúa ampliamente a través de los servicios de estadística nacionales e internacionales. En particular, los censos suministran información regular sobre la población.

Los métodos estadístico matemáticos emergieron desde la teoría de probabilidad, que data desde la correspondencia ciertamente entre Pierre de Fermat y Blaise Pascal (1654). Christian Huygens (1657) da el primer tratamiento científico que se conoce en la materia. El *Ars Conjectandi* (póstumo, 1713) de Jakob Bernoulli y la *Doctrina de Posibilidades* (1718) de Abraham de Moivre estudiaron la materia

como una rama de las matemáticas. En la era moderna, el trabajo de Kolmogorov ha sido un pilar en la formulación del modelo fundamental de la Teoría de Probabilidades, el cual es usado a través de la estadística. La teoría de errores se puede remontar a la *Opera Miscellanea* (póstuma, 1722) de Roger Cotes y al trabajo preparado por Thomas Simpson en 1755 (impreso en 1756) que aplica por primera vez la teoría de la discusión de errores de observación. La reimpresión (1757) de esta obra incluye el axioma de que errores positivos y negativos son igualmente probables y que hay unos ciertos límites asignables dentro de los cuales se encuentran todos los errores, se describen errores continuos y una curva de probabilidad.

Pierre-Simon Laplace (1774) hace el primer intento de deducir una regla para la combinación de observaciones desde los principios de la teoría de probabilidades. Laplace representó la ley de probabilidades de errores mediante una curva y dedujo una fórmula para la media de tres observaciones. También, en 1871, obtiene la fórmula para la ley de facilidad del error (término introducido por Lagrange, 1744) pero con ecuaciones inmanejables. Daniel Bernoulli (1778) introduce el principio del máximo producto de las probabilidades de un sistema de errores concurrentes. *El método de mínimos cuadrados*, el cual fue usado para minimizar los errores en mediciones, fue publicado independientemente por Adrien-Marie Legendre (1805), Robert Adrain (1808) y Carl Friedrich Gauss (1809). Gauss había usado el método en su famosa predicción de la localización del planeta enano Ceres en 1801. Pruebas adicionales fueron escritas por Laplace (1810, 1812), Gauss (1823), James Ivory (1825, 1826), Hagen (1837) , Friedrich Bessel (1838), WF Donkin (1844, 1856), John Herschel (1850) y Morgan Crofton (1870). Otros, Col van Ellis (1844), Augustus De Morgan (1864), Glaisher (1872) y Giovanni Schiaparelli (1875).

El siglo XIX incluye autores como Laplace, Silvestre Lacroix (1816), Littrow (1833), Richard Dedekind (1860), Helmert (1872), Hermann Laurent (1873), Liagre, Didion y Karl Pearson. Augustus De Morgan y George Boole mejoran la presentación de la teoría. Adolphe Quetelet (1796-1874) fue otro importante fundador de la estadística y quien introdujo la noción del «hombre promedio» (*l'homme moyen*) como un medio de entender los fenómenos sociales complejos como tasas de criminalidad, tasas de matrimonio o tasas de suicidios. Durante el siglo XX, la creación de instrumentos necesarios para asuntos de salud pública (epidemiología, estadística, etc.) y propósitos económicos y sociales (tasa de desempleo, econometría, etc.) necesitó de avances sustanciales en las prácticas estadísticas.

Hoy el uso de la estadística se ha extendido más allá de sus orígenes como un servicio al Estado o al gobierno. Personas y organizaciones usan la estadística para entender datos y tomar decisiones en ciencias naturales y sociales, medicina, negocios y otras áreas. La estadística es entendida generalmente no como un subárea de las matemáticas sino como una ciencia diferente «aliada». Muchas universidades tienen departamentos académicos de matemáticas y estadística separadamente. La estadística se enseña en departamentos tan diversos como psicología, educación y salud pública.

Al aplicar la estadística a un problema científico, industrial o social se comienza con un proceso o población a ser estudiado. Esta puede ser la población de un país, la de grandes cristalizados en una roca o la de bienes manufacturados por una fábrica en particular durante un período dado. También podría ser un proceso observado en varios instantes y los datos recogidos de esta manera constituyen una serie de tiempo.

Por razones prácticas, en lugar de compilar datos de una población entera, usualmente se estudia un subconjunto seleccionado de la población, llamado muestra. Datos sobre la muestra son recogidos de manera observacional o experimental. Los datos son entonces analizados estadísticamente lo cual sigue dos propósitos: descripción e inferencia.

El concepto matemático fundamental utilizado para entender la aleatoriedad es el de probabilidad. La estadística matemática (también llamada teoría estadística) es la rama de las matemáticas aplicadas que usa la teoría de probabilidades y el análisis matemático para examinar las bases teóricas de la estadística. El uso de cualquier método estadístico es válido solo cuando el sistema o población bajo consideración satisface los supuestos matemáticos del método. El mal uso de la estadística puede producir serios errores en la descripción e interpretación, afectando las políticas sociales, la práctica médica y la calidad de estructuras tales como puentes y plantas de reacción nuclear.

Incluso cuando la estadística es correctamente aplicada, los resultados pueden ser difícilmente interpretados por un no experto. Por ejemplo, el significado estadístico de una tendencia en los datos, que mide el grado en que la tendencia puede ser causada por una variación aleatoria en la muestra, puede no estar de acuerdo con el sentido intuitivo. El conjunto de habilidades estadísticas básicas (y el escepticismo) que una persona necesita para manejar información en el día a día se refiere como cultura estadística.

Este libro de problemas con ayudas es la primera parte de un conjunto de dos que comprenderá todas las fases del proceso estadístico. En este volumen se estudian mediante problemas los principales rasgos de la estadística descriptiva de una variable, de dos variables, los números índices y series temporales.

La novedad que presenta este manual es que todos los ejercicios tienen dos tipos de ayudas que aportan «pistas» de cómo resolver los ejercicios y los problemas. Así pues, el alumno puede consultarlas siempre que no sepa por dónde continuar mientras está resolviendo un ejercicio. De esta manera el estudiante evitará la desagradable sensación que una persona tiene cuando abandona la resolución de un ejercicio. Además, también se incluyen las soluciones completas de los ejercicios, muchos de ellos comentados con profundidad.

Es conveniente dejar claras dos cuestiones relevantes. La primera de ellas es que no hay que sacar la falsa idea de entender la estadística como una mera colección de métodos o técnicas útiles para el tratamiento de la información o, incluso lo que

es más, concluir que la estadística es lo que hacen los estadísticos. Aunque estas dos ideas no son desacertadas, tampoco permiten tener una visión completa de lo que es la estadística. La segunda es que nuestras decisiones se basan, cada vez más, en un flujo creciente de información que necesitamos sintetizar para evitar aquello de los árboles que impiden ver el bosque. Nuestras decisiones son de tipo condicionado, ya que las mismas se toman en función de algún tipo de información, tanto pasada como presente.

Este libro pretende ser un complemento didáctico de la teoría básica de estadística que se puede encontrar en otros numerosos libros que hoy en día se pueden encontrar en nuestras bibliotecas, así como sobre todo el manual *Introducción a la estadística aplicada a las ciencias sociales* de la Col·lecció Sapientia de Publicacions de la UJI, que puede considerarse el manual teórico que complementa este libro.

En definitiva, nuestra humilde pretensión es que este texto sirva de ayuda complementaria a todos aquellos estudiantes que se enfrentan (muchas veces con poco éxito) a la resolución de problemas de estadística descriptiva.

Los autores

Introducción

El presente libro de problemas se puede considerar como el primero de los dos complementos del manual *Introducción a la estadística aplicada a las ciencias sociales* de la Col·lecció Sapientia de Publicacions de la Universitat Jaume I, el cual consta fundamentalmente de contenidos teóricos, quedando el apartado de problemas en un segundo plano. Con este nuevo texto, basado casi exclusivamente en problemas resueltos, se completa parte del manual teórico y se facilita al estudiante una herramienta excelente para consolidar el aprendizaje de sus contenidos.

Los problemas cuentan con ayudas, siendo la última su resolución completa. Es decir, cada uno de los problemas tiene dos tipos de ayudas, que no son más que una breve información que puede facilitar al estudiante el arduo trabajo de resolver el problema. Las ayudas de tipo 1 son una mera orientación que tiene por objeto manifestar los contenidos que se deben consultar para poder resolver el problema. La ayuda de tipo 2 da bastante más información que la primera. Así, en muchas ayudas de este tipo se muestra parte de la resolución del ejercicio. Finalmente, en la resolución del problema se muestra con todo detalle los contenidos estadísticos que se utilizan y numerosos comentarios que permiten intuir la resolución del problemas similares.

Además, los problemas están clasificados por objetivos, ya que de esta manera el estudiante sabe en cada momento qué contenidos se están trabajando y, por tanto, puede consultar el manual teórico para revisar aquellas cuestiones en las que presente dificultades.

Por otra parte, este manual está dividido en cuatro unidades que hacen referencia a la estadítica descriptiva univariante, la estadística descriptiva bivariante, los números índices y, finalmente, las series temporales. Cada unidad está dividida en cuatro bloques: en el primero se proponen los enunciados de los problemas clasificados por objetivos. La segunda parte proporciona únicamente las ayudas de tipo 1 En el tercer bloque las ayudas son del tipo 2. El hecho de que para un mismo problema no se encuentren los dos tipos de ayudas conjuntamente tiene la pretensión de que el estudiante realice la consulta detallada de las ayudas, reforzando la idea de pensar antes de consultar. En la última parte se muestran las resoluciones completas de los problemas, las cuales están repletas de comentarios, gráficos y diagramas que facilitan su comprensión.

UNIDAD 1

Estadística descriptiva univariante

Introducción teórica

Como elementos introductorios de este capítulo, es conveniente recordar definiciones de elementos importantes, ya desarrolladas en diferentes materiales como los libros referenciados 1, 2 y 3, tales como:

Población: Es el conjunto de elementos, individuos o los sujetos a estudio y de los que se quiere obtener un resultado.

Parámetro: Es una medida descriptiva de la población total, de todas las observaciones.

Muestra: Conjunto de elementos que forman parte de la población total a la que representa.

Tamaño de la muestra: Es el número de elementos u observaciones que forman la muestra.

Estadístico: Es una medida descriptiva de la muestra y que estima el parámetro de la población.

Variables cualitativas y cuantitativas

Las variables en las que únicamente es posible un recuento del número de elementos de la población o muestra que poseen una de sus modalidades se llaman variables cualitativas o atributos (libros referenciados 4, 8, 14 y 19). Las modalidades de estos tipos de variables ni siquiera admiten una gradación y mucho menos una medida numérica. Son variables como el sexo de una persona, la confesionalidad, etc. Las modalidades que pueden tomar se denominan categorías. Así, las categorías de la variable sexo son masculino y femenino.

El resto de variables en las que, además de admitir el recuento del número de elementos de la población o muestra que poseen una de sus modalidades, también es posible asignarle una medida a la propia modalidad, se denominan variables cuantitativas. Son por ejemplo el peso, la altura, el sueldo mensual, el grado de dureza, etc.

Estas últimas variables, las cuantitativas, también pueden clasificarse en discretas y continuas. Una variable continua es aquella que puede tomar cualquier valor dentro de un rango dado. Independientemente de la proximidad de dos observaciones, si el instrumento de medida es suficientemente preciso, siempre se podrá encontrar una tercera observación entre las dos primeras.

Una variable discreta está limitada para ciertos valores, generalmente números enteros. Se diferencian de las continuas en que, dadas dos observaciones suficientemente

próximas, no se puede encontrar ninguna observación de la variable entre ellas. Son ejemplos el número de hijos de las familias, el número de vehículos que tienen las empresas, el número de turistas que visitan un país, etc.

La variable estadística se denota con mayúsculas. Asimismo, cada una de estas variables puede tomar distintos valores siendo su notación la siguiente:

$$X = (x_1, x_2, x_3, \ldots, x_{k-2}, x_{k-1}, x_k)$$

Tablas de frecuencia

Antes de construir las tablas de frecuencias, hay que realizar una serie de definiciones:

Se llama *frecuencia absoluta del valor* x_i al número de veces que aparece repetida la observación en la recopilación de datos. Se representa por n_i.

Se llama *frecuencia relativa del valor* x_i al cociente entre la frecuencia absoluta de x_i y el número total de datos n. Se representa por f_i y, evidentemente, es la proporción en que se encuentra el valor x_i dentro del conjunto de datos en tanto por uno; $f_i = \frac{n_i}{n}$.

Por otra parte, suponiendo que se dispondrá de *k* datos diferentes, se cumple que la suma de todos los n_i es n ($n_1 + n_2 + ... + n_k = n$), y también que la suma de las frecuencias relativas es igual a la unidad ($f_1 + f_2 + ... + f_k = 1$).

Se llama *frecuencia absoluta acumulada del valor* x_i al número de datos de la recopilación que son menores o iguales que x_i. Se representa por N_i y su valor se calcula a partir de las frecuencias absolutas; $N_i = n_1 + n_2 + ... + n_i$ (asumiendo que $x_1 < x_2 < ..... < x_i$).

Se llama *frecuencia relativa acumulada del valor* x_i al cociente entre la frecuencia absoluta acumulada de x_i y el número total de datos n. Se representa por F_i y, evidentemente, es la proporción en que se encuentran los valores menores o iguales a x_i dentro del conjunto de datos en tanto por uno; $F_i = \frac{N_i}{n}$. También hay otra manera de calcular F_i a partir de las frecuencias relativas, pues $F_i = f_1 + f_2 + ... + f_i$. (asumiendo que $x_1 < x_2 < ..... < x_i$).

Las frecuencias acumuladas también cumplen dos propiedades triviales como consecuencia de sus definiciones: suponiendo que se dispusiera de k datos diferentes, se cumple que $N_k = n$ y $F_k = 1$.

Es importante remarcar que para calcular frecuencias acumuladas es necesario que las variables por estudiar sean *ordenables*, es decir, debe ser posible establecer una relación de orden entre las variables. En otros casos, no tiene ningún sentido realizar estos cálculos.

Estas definiciones permiten resumir los datos. Sin embargo, la manera más adecuada para sintetizar los datos es mediante lo que se denomina tabla de frecuencias. En ella aparecen distribuidas los datos según las frecuencias. Al mismo tiempo refleja todos los conceptos mencionados con anterioridad.

En ocasiones el número de datos diferentes que se está estudiando es muy numeroso. Entonces, si se decidiera construir una tabla como la anterior, la columna relativa a las x_i sería muy extensa, únicamente hay que pensar en doscientos datos diferentes dentro de una recopilación de cuatrocientos.

La solución a esta cuestión consiste en agrupar los datos en intervalos o clases, de modo que cada dato pertenezca a uno y solo un intervalo. En consecuencia, los conceptos relativos a la frecuencia que hasta ahora se referían a los valores diferentes de los datos, al realizar la agrupación, deben hacer referencia a los intervalos.

Esta práctica, a pesar de que ayuda a resumir y clarificar la información, tiene en cambio un inconveniente: se pierde información sobre la propia distribución de datos. Al agruparlas en los intervalos los valores reales se «difuminan».

Un intervalo se suele representar por $[Li_{-1}, L_i)$ y se define como el conjunto formado por todos los valores reales que son mayores o iguales que Li_{-1} (*Extremo inferior*) y menores que L_i (*Extremo superior*).

Se llama marca de clase a la media aritmética de los dos extremos del intervalo. Es evidentemente el valor central del intervalo ya que equidista de los extremos. Se denota por c_i . Se calcula $c_i = \frac{L_{i-1} + L_i}{2}$.

Se llama amplitud de un intervalo a la distancia que hay entre los extremos. Se denota por a_i y se calcula $a_i = L_i - L_{i-1}$.

Se llama densidad de frecuencia absoluta de un intervalo al cociente entre la frecuencia absoluta del intervalo y su amplitud. Se denota por d_i. Se calcula $d_i = \frac{n_i}{a_i}$.

Sin embargo, en la literatura matemática es posible encontrar varias reglas para calcular el número adecuado de intervalos a partir del número de datos, como que no puede superar el 10 % del número total de datos o como el método de la raíz. Según este método el número de clases es igual a la raíz cuadrada del número de datos:

$$\text{Número clases} = \sqrt{\textit{número de datos}}$$

Se llama recorrido de un conjunto de datos a la diferencia entre el valor más grande y el más pequeño del conjunto. Se denota por *Re*.

En consecuencia, para averiguar la amplitud de la clase calculamos:

$$\text{Amplitud} = \frac{recorrido}{número\ de\ clases}$$

Conociendo el número de intervalos y la amplitud se pueden construir fácilmente todos los intervalos. Al finalizar la construcción de todos los intervalos es necesario comprobar que todos los datos pertenecen a un y solo un intervalo. Si no es así, hay que realizar alguna modificación en la amplitud o en el número de intervalos.

Gráficos estadísticos

Los gráficos también son muy útiles para describir los conjuntos de datos (referencias 15, 20 y 23). De hecho, un gráfico estadístico permite formarse una primera idea de la distribución de los datos tan solo con una observación. No obstante, hay que tener cuidado pues en algunas ocasiones los gráficos presentan «tendencias» no atribuibles al quehacer matemático.

Diagrama de sectores o diagrama circular: Es un círculo dividido en diferentes sectores. El área de cada sector es proporcional a la frecuencia que se quiera representar, sea absoluta o relativa.

Para calcular el ángulo asociado a cada frecuencia se aplica una simple proporción: el ángulo asociado a una frecuencia absoluta n_i es igual a $f_i \cdot 360º$ ($f_i = \frac{n_i}{n}$). Para la frecuencia absoluta acumulada se razona de la misma manera.

Diagrama de barras: Se utiliza para representar los datos que no están agrupados. Consiste en colocar sobre un eje horizontal los distintos valores que toma la variable estadística, y sobre cada uno de ellos levantar un rectángulo de altura igual a la frecuencia (del tipo que se esté representando). Todos los rectángulos deben tener la misma amplitud.

Histogramas: Se utilizan para representar datos agrupados en intervalos. Consiste en colocar sobre un eje horizontal los diferentes intervalos. Sobre cada uno de ellos se construye un rectángulo de *superficie* igual a la frecuencia que se esté representando. Así, las alturas de los rectángulos deben ser las densidades de los intervalos. Hay que notar que en el eje horizontal aparecen reflejadas las marcas de las clase.

Polígono de frecuencias: Es menos utilizado que los diagramas de barras y los histogramas, pero pueden sustituirlos. Consiste en unir mediante líneas poligonales los extremos superiores de las barras si se trata de datos sin agrupar, o el punto medio de la base superior de los rectángulos, si se trata de histogramas.

Pictograma: Se suele utilizar para expresar un atributo. Se suelen utilizar iconos que se identifican con la variable (ejemplo una bombilla, si la variable es la energía electrica consumida en un hogar) y su tamaño es proporcional a la frecuencia.

Medidas de posición

Son coeficientes que tratan de representar una determinada distribución; pueden ser de dos tipos, centrales y no centrales.

Medidas Centrales

Media aritmética

Es el valor que habitualmente se toma como representación de los datos. Es la suma de todos los valores de la variable dividida entre el número total de elementos. Si los datos están agrupados, se toma la marca de la clase como representante del intervalo y se realizan todos los cálculos como si los valores de la variable fueran las marcas de las clases.

Si se considera una variable estadística X que tiene k valores diferentes, que se representan por x_i y sus frecuencias para n_i, entonces la media aritmética se calcula:

$$\text{Media aritmética: } \frac{n_1x_1 + n_2x_2 + \ldots\ldots + n_kx_k}{n}$$

La media aritmética cumple las siguientes propiedades:

- La suma de las desviaciones de los valores de la variable respecto a la media aritmética es 0.
- Si a todos los valores de la variable se les suma una misma constante, la media aritmética queda aumentada en dicha constante.
- Si todos los valores de la variable se multiplican por una misma constante la media aritmética queda multiplicada por dicha constante.
- Si una variable Y es transformación lineal de otra variable X ($Y = a \cdot X + b$; a y b números reales), la media aritmética de Y sigue la misma transformación lineal respecto a la media aritmética de X. Es decir: $\overline{Y} = a \cdot \overline{X} + b$.
- Si en un conjunto de valores se pueden obtener 2 o más subconjuntos disjuntos que suponen una partición del conjunto total de valores, la media aritmética del conjunto se relaciona con la media aritmética de cada uno de los subconjuntos disjuntos de la siguiente forma: $\overline{X} = \frac{\sum \overline{X}_i \cdot N_i}{n}$ (siendo $\overline{X}_i$ la media de cada subconjunto y N_i el número de elementos de cada subconjunto).

Media aritmética ponderada

A veces, no todos los valores de la variable tienen el mismo peso. Es decir, cada uno de los valores que toma la variable tiene asignado un número que indica su importancia, el cual es independiente de la propia frecuencia absoluta.

El cálculo de la media aritmética ponderada en estas casos sigue la siguiente expresión, donde w_i es el peso asociado a cada valor de la variable x_i.

$$\overline{X}_W = \frac{\sum_{i=1}^{k} x_i w_i n_i}{\sum_{i=1}^{k} w_i n_i}$$

Media geométrica

Puede utilizarse para mostrar cambios porcentuales en una serie de números positivos. Por lo tanto, tiene una amplia aplicación en los negocios y en la economía. La media geométrica proporciona una medida precisa de un cambio porcentual medio en una serie de números. Se representa por G y su cálculo –efectuando la notación habitual– sigue la siguiente expresión.

$$G = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot ...... x_k^{n_k} \cdot}$$

Utilizando la notación potencial, también se puede presentar por:

$$G = (x_1^{n_1} \cdot x_2^{n_2} \cdot ...... x_k^{n_k})^{\frac{1}{n}}$$

Media harmónica

Se representa por H y es la inversa de la media aritmética de las inversas de los valores de la variable, con expresión:

$$H = \frac{n}{\sum \frac{n_i}{x_i}} = \frac{n}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + ... + \frac{n_k}{x_k}}$$

$n =$ *número total de datos*

$x_i =$ *valores diferentes que toman la variable*

$n_i =$ *frecuencia absoluta de* x_i

Se utiliza para calcular el valor medio de magnitudes expresadas en términos relativos como velocidades, tiempos, rendimiento, tipo de cambio monetario, etc. Su principal contrariedad es que cuando algún valor de la variable es 0 o próximo a cero no se puede calcular.

En muchas ocasiones, no es necesario aplicar la fórmula anterior. Únicamente hay que tener presente el concepto de media aritmética.

Mediana

La mediana es el valor de la variable que divide las observaciones en dos grupos de igual número de elementos, de modo que en el primer grupo todos los datos sean menores o iguales que la mediana, y en el otro grupo, todas los datos sean mayores o iguales. Por lo tanto, es una cantidad que indica orden dentro de la ordenación.

Datos no agrupados

Al ordenar los datos, la posición que ocupa la mediana se determina dividiendo el número total de valores entre 2 ($\frac{n}{2}$) o lo que es lo mismo, calculando el 50 % del total de datos (0,5 · n). Hay que tener en cuenta, sin embargo, la paridad de *n*:

- Cuando haya un número impar de valores, la mediana será justo el valor central. Si hay muchos datos el cálculo no es inmediato, hay que construir la tabla de frecuencias y fijarse en la columna de las frecuencias absolutas acumuladas N_i. La mediana será el valor de variable que tenga la frecuencia absoluta acumulada igual a $\frac{n}{2}$. Es decir:

$$\text{si}\, N_{i-1} \leq \frac{n}{2} \leq N_i \rightarrow Me = x_i$$

- Cuando haya un número par de valores, la mediana será la media aritmética de los dos valores centrales de la variable. Del mismo modo que en el caso anterior, si el conjunto de observaciones es numeroso, es necesario construir la tabla de frecuencias y fijarse en la columna de las N_i. Si al calcular $\frac{n}{2}$ este resulta ser un valor menor que una frecuencia absoluta acumulada, la mediana se calculará de la misma manera que en el caso anterior; es decir, si $N_{i-1} \leq \frac{n}{2} \leq N_i \rightarrow Me = x_i$. Sin embargo, si coincide $\frac{n}{2}$ con algun N_i, para obtenerla se realizará el cálculo siguiente: $Me = \frac{x_i + x_{i+1}}{2}$. Los ejemplos siguientes clarifican los cálculos.

Datos agrupados

En distribuciones agrupadas es necesario determinar el intervalo [Li-1, Li) en el que se encuentra la mediana. Este intervalo se determina siguiendo exactamente los mismos procedimientos mencionados en el apartado anterior; se realiza el mismo que en el caso de datos no agrupados. La diferencia radica en que se obtendrá un intervalo en lugar de un valor. Una vez se tiene el intervalo $[L_{i-1}, L_i)$, la mediana se calcula:

$$Me = L_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} a_i \quad \text{donde,}$$

L_{i-1} Límite inferior
N_{i-1} Es la frecuencia absoluta acumulada de la clase «anterior» a la clase mediana
n_i Es la frecuencia de la clase mediana
a_i Es la amplitud de la clase mediana

Es evidente que lo que se pretende es calcular un representante del intervalo con el objeto de fijar la mediana en un valor. Una posibilidad hubiera sido considerar la marca de clase, sin embargo, el criterio usualmente más seguido no es este sino el de la fórmula antes mencionada.

En esta fórmula en primer lugar se considera el supuesto de que los datos están uniformemente distribuidos dentro de cada intervalo. Teniendo este hecho en cuenta, se puede observar que la fórmula es una relación de proporcionalidad entre las posiciones que ocupan los valores de la variable y la amplitud de los intervalos.

Moda

Es el valor de la variable que más veces se repite, es decir, el valor que tiene mayor frecuencia absoluta.

Pueden existir distribuciones con más de una moda: bimodales, trimodal, etc.

Datos no agrupados

En las distribuciones sin agrupar, la obtención de la moda es inmediata.

Datos agrupados

En los supuestos que la distribución venga dada en intervalos, se pueden producir dos casos: que tengan la misma amplitud, o que esta sea distinta. En ambos casos el objetivo es encontrar un valor que represente la moda.

Intervalos con la misma amplitud

Es evidente que una vez determinada la mayor frecuencia a esta no le corresponde un valor sino un intervalo. Entonces no tendremos un valor modal sino un intervalo modal. Para calcular el representado del intervalo que haga el papel de moda hay distintos criterios. En el texto se recoge el siguiente. En primer lugar se calcula el intervalo donde se encuentra la moda, es decir, el intervalo modal $[L_{i-1}, L_i)$, el cual tiene mayor frecuencia absoluta (n_i). Posteriormente se calcula la moda de la siguiente manera:

$$Mo = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} \cdot a_i$$

Donde:

L_{i-1}: extremo inferior del intervalo modal
a_i: amplitud del intervalo
n_{i-1}, n_{i+1}: frecuencias de los intervalos anteriores y posterior respectivamente del intervalo modal

Del mismo modo que la mediana, la fórmula tiene el supuesto de que los datos están uniformemente repartidas dentro de cada intervalo. Además, siguiendo este criterio se puede observar que la moda estará más cerca de aquel intervalo adyacente con mayor frecuencia absoluta.

Medidas no Centrales

Percentiles o cuantiles

Son medidas de localización similares a la mediana. Su función es informar del valor de la variable que ocupará la posición (en tanto por ciento) que nos interese respecto de todo el conjunto de observaciones.

Podemos decir que los cuantiles son unas medidas de posición que dividen la distribución en un cierto número de partes.

Las más importantes son:

- *Cuartiles*, dividen la distribución en cuatro partes iguales (tres divisiones). C_1, C_2, C_3, correspondientes a 25 %, 50 %, 75 %. Por ejemplo, el 1.º cuartil tiene un 25 % de los datos menores o iguales a él, el segundo cuartil es la mediana, etc.

- *Deciles*, dividen la distribución en 10 partes iguales (9 divisiones). D_1,..., D_9, correspondientes a 10 %,..., 90 %.

- *Percentiles*, dividen a la distribución en 100 partes (99 divisiones). P_1,..., P_{99}, correspondientes a 1 %,..., 99 %. Por ejemplo, el valor correspondiente al percentil 65, tiene un 65 % de los datos menores o iguales a él.

Hay un valor en el que coinciden los cuartiles, los deciles y percentiles. Es la mediana, ya que: $P_{50} = C_2 = D_5$. El cálculo de los cuantiles sigue el mismo procedimiento que el que se ha utilizado en la mediana, tanto para los datos agrupados como para los datos sin agrupar. Así, en general se calcula la posición en que se encuentra el cuantil y después se calcula. Se distingue entre distribuciones agrupadas y las que no lo están:

Datos no agrupados

Primero se calcula la posición que ocupa el cuantil que se está estimando. Así, si Q_a representa el cuantil que deja por debajo de él un *a* (%) de los datos:

$$si\ N_{i-1} \leq \frac{a}{100} \cdot n \leq N_i \rightarrow Q_a = x_i$$

$$\text{en el supuesto que } \frac{a}{100} \cdot n = N_i \rightarrow Q = \frac{x_i + x_{i+1}}{2}$$

Datos agrupados

En distribuciones agrupadas, es necesario determinar el intervalo [Li-1, Li) en el que se encuentra el cuantil. Este intervalo se determina siguiendo exactamente los mismos procedimientos mencionados en el apartado anterior; se realiza el mismo que en el caso de datos no agrupados. La diferencia radica en que se obtendrá un intervalo en lugar de un valor. Un vez se tiene el intervalo [Li-1, Li), el cuantil se calcula:

$$Me = L_{i-1} + \frac{\frac{a}{100} \cdot n - N_{i-1}}{n_i} a_i \text{ donde,}$$

L_{i-1} Límite inferior de la clase mediana
N_{i-1} Es la frecuencia absoluta acumulada de la clase «anterior» a la clase mediana
n_i Es la frecuencia de la clase mediana
a_i Es la amplitud de la clase mediana

Medidas de dispersión

Son complementarias de las de posición, en el sentido que señalan la dispersión del conjunto de todos los datos de la distribución, respecto de la medida o medidas de localización adoptadas.

Recorrido

Se define como la diferencia entre el mayor y menor valor de las variables de una distribución de datos, es decir:

$$\text{Re} = \max(x_i) - \min(x_i)$$

Recorrido intercuartílico

Se define como la distancia que hay entre el tercer y el primer cuartil, es decir:

$$\text{Re} = C_3 - C_1$$

Desviación media respecto de la mediana

Se define como la media aritmética de los valores absolutos de las desviaciones de los valores de la variable respecto de la mediana. Responde a la siguiente expresión:

$$D_{|Me|} = \frac{\sum_{i=1}^{k} |x_i - Me| \cdot n_i}{n}$$

Varianza

Se define como la media aritmética de los cuadrados de las desviaciones de los valores de la variable respecto de la media aritmética de la distribución. Responde a la expresión:

$$s^2 = \frac{(x_1 - \overline{X})^2 \cdot n_1 + (x_2 - \overline{X})^2 \cdot n_2 + \ldots + (x_k - \overline{X})^2 \cdot n_k}{n} = \frac{\sum_{i=1}^{k} (x_i - \overline{X})^2 \cdot n_i}{n}$$

Como se puede observar en la definición, la varianza es un promedio del cuadrado de los errores que se cometen al considerar la media aritmética como «el representante» de todos y cada uno de los datos.

Por otra parte, una de las principales dificultades que presenta la varianza es la unidad, ya que viene dada en unidades al cuadrado (h^2, m^2, etc.). La manera de solucionar esta circunstancia es estimando la raíz cuadrada.

Desviación típica o desviación estándar

Se define como la raíz cuadrada, con signo positivo, de la varianza. Responde a la siguiente expresión:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{k} (x_i - \overline{X})^2 \cdot n_i}{n}}$$

En las definiciones anteriores se han estado considerando datos no agrupados. Si lo fueran, únicamente hay que emplear las marcas de clases como representantes de los intervalos. Es decir, $c_i = x_i$.

Por otra parte, se pueden definir dos estadísticos de dispersión más, llamados quasivariancia y cuasidesviación típica como:

$$s_{n-1}^2 = \frac{n}{n-1} s^2 \quad \text{y} \quad s_{n-1} = \sqrt{\frac{n}{n-1}} \cdot s$$

Estos estadísticos tienen mucho interés en la Estadística Inferencial como se verá en capítulos posteriores.

La varianza cumple las siguientes propiedades:

- La varianza es siempre un valor no negativo o 0. Únicamente puede ser 0 si todos los datos son iguales. En este caso es evidente que $\overline{X} = x_i$ para todo los posibles valor del índice.

- Si a todos los valores de la variable se les suma una constante la varianza no se modifica.

- Si todos los valores de la variable se multiplican por una constante la varianza queda multiplicada por el cuadrado de dicha constante.

- Si una variable X' es transformación lineal de otra variable X ($X' = a \cdot X + b$; a y b números reales), la varianza de X' se obtiene a partir de la de X del modo $s'^2 = a^2 \cdot s^2$.

Las medidas de dispersión absolutas son unos indicadores que presentan dificultades a la hora de comparar la representatividad de las medidas de tendencia central entre dos distrubucions de datos diferentes. Por ello, a veces se recurre a medidas de dispersión relativas.

El coeficiente de variación de Pearson

Es una de las más significativas y determina el grado de significación de un conjunto de datos relativo a su media aritmética. Se define como el cociente entre la desviación típica y la media aritmética de la distribución de datos.

$$V_X = \frac{s}{|\overline{X}|}$$

MEDIDAS DE FORMA

Nos dan información de la forma del histograma, de su simetría y de la menor o menor proximidad de los valores de la variable respecto de su promedio.

Coeficiente de asimetría de Fisher

Las medidas de asimetría permiten determinar, sin que sea necesario hacer las representaciones gráficas, el grado de simetría que presentan los datos respecto a un valor central de la variable estadística, normalmente la media aritmética. Por tanto, esta medida debe reflejar dos aspectos: la distancia de cada observación respecto a la media aritmética (es decir, la diferencia entre cada valor y la media

aritmética: $x_i - \overline{x}$) y la frecuencia de cada una de estas distancias (la que coincidirá, evidentemente, con la frecuencia de cada observación). De esta manera, intuitivamente, si «predominan» las distancias negativas sobre las positivas (por ser más frecuentes o ser distancias muy grandes), entonces la distribución es asimétrica a izquierdas. Si por el contrario, se da la situación opuesta entonces la distribución es asimétrica a derechas. Para finalizar, si las distancias negativas y las positivas se «compensan», entonces la distribución es simétrica.

Ahora pues, lo que hay que encontrar es el estadístico que determine la asimetría de la distribución de datos. Como la asimetría está directamente relacionada con las desviaciones respecto a la media aritmética, una primera aproximación puede ser la media de las desviaciones, es decir, $\frac{\sum_{i=1}^{k}(x_i - \overline{X})n_i}{n}$. Sin embargo, ya es conocido que esta suma es cero (propiedades de la media aritmética).

Por otra parte, como lo que nos interesa es conocer el signo de las desviaciones, tampoco podemos emplear el cuadrado de las desviaciones. Así pues, parece coherente tomar una potencia de grado tres de las desviaciones y calcular la media. Así, si llamamos

$$\text{m} = \frac{\sum_{i=1}^{k}(x_i - \overline{X})^3 n_i}{n},$$

por lo que se cumple:

$$\begin{array}{ll} si \;\; m = 0 & la \; distribución \;\; es \; simétrica \\ si \;\; m > 0 & la \; distribución \; es \;\; asimétrica \;\; positiva \\ si \;\; m < 0 & la \; distribución \; es \;\; asimétrica \;\; negativa \end{array}$$

De esta manera se obtiene el *coeficiente de asimetría de Fisher*.

$$g_1 = \frac{m}{s^3} = \frac{\frac{\sum_{i=1}^{k}(x_i - \overline{X})^3 n_i}{n}}{\left(\sqrt{\frac{\sum_{i=1}^{k}(x_i - \overline{X})^2 n_i}{n}}\right)^3}$$

Índice

Curtosis

Para estudiar el grado de curtosis de una distribución hay que tomar un modelo teórico como referencia, la representación gráfica tenga forma de campana simétrica. No es extraño pues, que se tome el modelo normal, ya que, como ya se ha mencionado con anterioridad, se puede decir que es el modelo campaniforme por antonomasia.

De esta manera, tomando este modelo como referencia, se dice que una distribución es *leptocúrtica* si es más apuntada que la distribución normal. Si es menos apuntada se le llama *platicúrtica*. Finalmente, si tiene el mismo apuntamiento que una distribución normal se le llama *mesocúrtica*.

Del mismo modo que en el caso del estudio de la asimetría, hay un coeficiente que permite clasificar los datos según la curtosis. En este caso, el coeficiente no es tan intuitivo, por lo que únicamente se dará la definición y su interpetación. Como en el caso de la otra medida de forma, este indicador tampoco tiene dimensión.

$$g_2 = \frac{\dfrac{\sum_{i=1}^{k}(x_i - \overline{X})^4 n_i}{n}}{\left(\dfrac{\sum_{i=1}^{k}(x_i - \overline{X})^2 n_i}{n}\right)^2} - 3$$

La idea del apuntamiento de una distribución de datos sale de la comparación de la frecuencia de los valores centrales de una distribución con la frecuencia de los valores centrales en un modelo teórico normal que tenga la misma media y la misma desviación típica que la distribución que se está estudiando.

Como en un modelo normal se cumple que $\dfrac{\dfrac{\sum_{i=1}^{k}(x_i - \overline{X})^4 n_i}{n}}{s^4} = 3$, entonces:

Una distribución será:

mesocúrtica (normal)	si $g_2 = 0$
leptocúrtica	si $g_2 > 0$
platicúrtica	si $g_2 < 0$

Por último, debemos remarcar que el estudio de la curtosis no implica necesariamente que las distribuciones sean simétricas. Así, por ejemplo, nos podríamos encontrar distribuciones de observaciones que sean leptocúrticas y, al mismo tiempo, asimétricas positivas.

Cajas y bigotes (Box-plot)

Un diagrama de cajas y bigote (conocido también como *Box and whisker plot* en inglés), es una representación gráfica de los datos que permite determinar con mucha facilidad y de una manera visual la tendencia central, la variabilidad, la asimetría y la existencia de valores anómalos de un conjunto de observaciones (*outliers*). De alguna manera, se puede decir que es uno de los gráficos que más y mejor resumen los conjuntos de datos.

El diagrama de cajas emplea el resumen de los 5 números: la menor observación, la mayor observación, el primer cuartil, la mediana y el tercer cuartil.

Medidas de concentración

Estudian el grado de concentración de una magnitud, normalmente económica, en determinados individuos. En cierto modo es un término opuesto a la equidad en el reparto. Se denomina *concentración* al grado de equidad en el reparto de la suma total de los valores de la variable considerada (renta, salarios, etc.).

Las infinitas posibilidades que pueden adoptar los valores se encuentran entre los dos extremos:

Concentración máxima, cuando un solo individuo percibe el total y los demás nada; en este caso, se está ante un reparto no equitativo:

$$\text{el que recibe } x_1 = \text{el que recibe } x_2 = \ldots\ldots = \text{el que recibe } x_{k-1} = 0 \text{ y el que recibe } x_k = \text{el total}$$

Concentración mínima, cuando el conjunto total de valores de la variable esta repartido por igual, en este caso se está ante un reparto equitativo:

$$\text{el que recibe } x_1 = \text{el que recibe } x_2 = \ldots\ldots = \text{el que recibe } x_{k-1} = \text{el que recibe } x_k$$

Hay diferentes medidas de concentración, pero en el texto se va a estudiar el índice de Gini; por ser un coeficiente, será un valor numérico. Para obtenerlo es necesario realizar un conjunto de cálculos.

Se supone que hay una distribución de rentas ($x_i \cdot n_i$) donde *i* toma los valores de 1 hasta *k* (por ejemplo, x_i son los sueldos y *n* el número de personas que cobran ese sueldo) de la que se formará una tabla con las columnas siguientes:

1) Los productos $x_i \cdot n_i$ indicarán la renta total percibida por los n_i rentistas de renta individual x_i.
2) Las frecuencias absolutas acumuladas N_i.
3) Los totales acumulados u_i que se calculan de la siguiente forma:

$$u_1 = x_1 n_1$$
$$u_2 = x_1 n_1 + x_2 n_2$$
$$u_3 = x_1 n_1 + x_2 n_2 + x_3 n_3$$
$$u_4 = x_1 n_1 + x_2 n_2 + x_3 n_3 + x_4 n_4$$
....
$$u_k = x_1 n_1 + x_2 n_2 + x_3 n_3 + x_4 n_4 + \ldots\ldots\ldots\ldots + x_k n_k$$

Por tanto, se puede decir que:

$u_j = \sum_{i=1}^{j} x_i \cdot n_i$ para cualquier valor de j desde 1 hasta k.

4) La columna total de frecuencias acumuladas relativas, que se expresa en tanto por ciento y que se representa por p_i, vendrá dado por la siguiente notación:

$$p_i = \frac{N_i}{n}$$

5) La columna de renta acumulada relativa, que se expresa en tanto por ciento y que se representa por la expresión:

$$q_i = \frac{u_i}{u_k}$$

Por tanto, ya se puede confeccionar la tabla:

x_i	n_i	$x_i n_i$	N_i	u_i	$p_i = \frac{N_i}{n}$	$q_i = \frac{u_i}{u_k}$	pi - qi
x_1	n_1	$x_1 n_1$	N_1	u_1	p_1	q_1	$p_1 - q_1$
x_2	n_2	$x_2 n_2$	N_2	u_2	p_2	q_2	$p_2 - q_2$
...	...	...	...	...	...	...	...
x_k	n_k	$x_k n_k$	N_k	u_k	100	100	0

Como se puede ver, la última columna es la diferencia entre las dos penúltimas; esta diferencia sería 0 para la concentración mínima en la que se cumple $p_i = q_i$ para cualquier i, por tanto su diferencia sería cero.

Analíticamente el índice de Gini: $I_G = \dfrac{\sum_{j=1}^{k-1}(p_i - q_i)}{\sum_{j=1}^{k-1} p_i}$

Este índice tendrá los valores:

- $i_G = 0$ cuando $p_i = q_i$ concentración mínima
- $i_G = 1$ cuando $q_i = 0$ concentración máxima

Por otra parte, si se representan gráficamente los q_i en el eje vertical y los p_i en la horizontal se obtendrá la curva de concentración o curva de Lorenz. Se puede comprobar que esta curva resultante siempre aparecerá «por debajo» de la diagonal del primer cuadrante, la cual representa la concentración mínima. Además, cuando más se aproxime esta curva a la diagonal, menor será la concentración.

A continuación, se desarrollará los objetivos y los ejercicios correspondientes a este capítulo. Cabe recordar que el material desarrollado y el resultado de algunos ejercicios son aplicaciones desarrolladas con el software R (referencias bibliográficas 13, 18 y 22).

Objetivos

Los problemas deben permitir que los alumnos alcanzan los objetivos didácticos:

1*a*) Conocer los conceptos básicos de las variables estadísticas.
1*b*) Saber clasificar las variables estadísticas.
1*c*) Saber analizar y realizar tablas de frecuencias de un conjunto de datos.
1*d*) Conocer las diferencias entre las tablas de datos sin agrupar y las tablas de datos agrupados.
1*e*) Saber interpretar y construir los principales gráficos estadísticos.
1*f*) Conocer los conceptos y saber realizar los cálculos de las medidas de tendencia central y de dispersión. Concretar con la aplicación del coeficiente de variación de Pearson en aquellas situaciones que lo requieran.
1*g*) Conocer los principales estadísticos que miden la forma de los datos a partir de los gráficos.
1*h*) Saber calcular e interpretar el índice de Gini, así como saber realizar la curva de Lorenz para medir la equidad de un reparto.

La tabla siguiente nos muestra cómo están distribuidos los objetivos según los ejercicios:

Objetivos Ejercicio	***1a***	***1b***	***1c***	***1d***	***1e***	***1f***	***1g***	***1h***
1	**x**	**x**						
2	x	x			x			
3	x	x			x			
4		x	x	x				
5		x	x	x				
6		x	x		x	x		
7		x					x	
8								x
9	x				x	x		

Enunciados

1*a*) Conocer los conceptos básicos de las variables estadísticas.
1*b*) Saber clasificar las variables estadísticas.

Ejercicio 1

Clasifica las siguientes variables, justificando el por qué de la elección:

a) Color de los coches.
b) Marcas de ordenadores.
c) Longitud de carreteras en metros.
d) Nivel de estudios.
e) Número de hijos de una familia.
f) Número de alumnos de estadística en una carrera.
g) Metros de altitud de las montañas.
h) Profesiones de las personas.
i) Sueldo mensual de los trabajadores de las empresas del sector cerámico.

1*a*) Conocer los conceptos básicos de las variables estadísticas.
1*b*) Saber clasificar las variables estadísticas.
1*e*) Saber interpretar y construir los principales gráficos estadísticos.

Ejercicio 2

Actualmente, se está estudiando en las distintas comunidades autónomas el número de hijos por familia para estudiar la natalidad. Uno de los trabajadores que está haciendo las encuestas, recoge los datos de su barrio donde hay 100 familias. Ha obtenido los siguientes datos que aparecen en la tabla sguiente:

1	3	3	0	4	3	1	4	0	0
2	1	0	3	1	2	1	4	1	2
3	3	4	2	0	4	3	0	2	3
1	3	4	2	2	4	4	4	2	1
4	2	1	1	0	1	1	2	3	0
3	3	3	1	1	3	3	0	2	3
4	3	0	3	1	2	2	1	2	3
3	2	1	3	1	3	4	4	4	1
3	0	3	1	0	4	3	2	3	2
1	2	0	2	0	0	2	2	3	4

a) Construye el gráfico que consideres más adecuado con las frecuencias acumuladas.
b) Construye el polígono de frecuencias con las frecuencias acumuladas.

1*a*) Conocer los conceptos básicos de las variables estadísticas.
1*b*) Saber clasificar las variables estadísticas.
1*e*) Saber interpretar y construir los principales gráficos estadísticos.

Ejercicio 3

Los sueldos, en miles de euros mensuales de 40 empresarios del sector de la construcción del año 2007 son:

3,9	4,7	3,7	5,6	4,3	4,9	5,0	6,1	5,1	4,5
5,3	3,9	4,3	5,0	6,0	4,7	5,1	4,2	4,4	5,8
3,3	4,3	4,1	5,8	4,4	4,8	6,1	4,3	5,3	4,5
4,0	5,4	3,9	4,7	3,3	4,5	4,7	4,2	4,5	4,8

Se quiere estudiar si realmente son bastante altos y cuál es su distribución. Para conseguirlo:

a) Representa gráficamente la información recogida.
b) Crea la misma representación en 4 clases para poder diferenciar de forma más clara los tipos de sueldos.

1*b*) Saber clasificar las variables estadísticas.
1*c*) Saber analizar y realizar tablas de frecuencias de un conjunto de datos.
1*d*) Conocer las diferencias entre las tablas de datos sin agrupar y las tablas de datos agrupados.

Ejercicio 4

La recopilación de 20 datos correspondientes al número de llamadas de teléfono registradas en una empresa durante los días de preparación de material para una feria de muestras durante el período de 9 a 12 horas.

15,5, 10, 5, 5, 6, 5, 6, 5, 6, 7, 10, 10, 12, 11, 11, 12, 15, 12, 15

Se quiere estudiar si realmente hay variación a lo largo de los días de las llamadas que se reciben. Por este motivo se pide confeccionar una tabla de frecuencias que recoja esta información.

1*b*) Saber clasificar las variables estadísticas.
1*c*) Saber analizar y realizar tablas de frecuencias de un conjunto de datos.
1*d*) Conocer las diferencias entre las tablas de datos sin agrupar y las tablas de datos agrupados.

Ejercicio 5

Una empresa está haciendo el estudio del dinero que se gasta la gente para comprar una segunda casa como complemento de la primera vivienda. Reducir los datos de los euros y en número de familias que han comprado este tipo de vivienda. A continuación se puede ver los datos:

Euros	Familias
0-50000	2145
50000-75000	1520
75000-100000	840
100000-115000	955
115000-135000	1110
135000-140000	2342
140000-150000	610
150000-200000	328
>200000	150

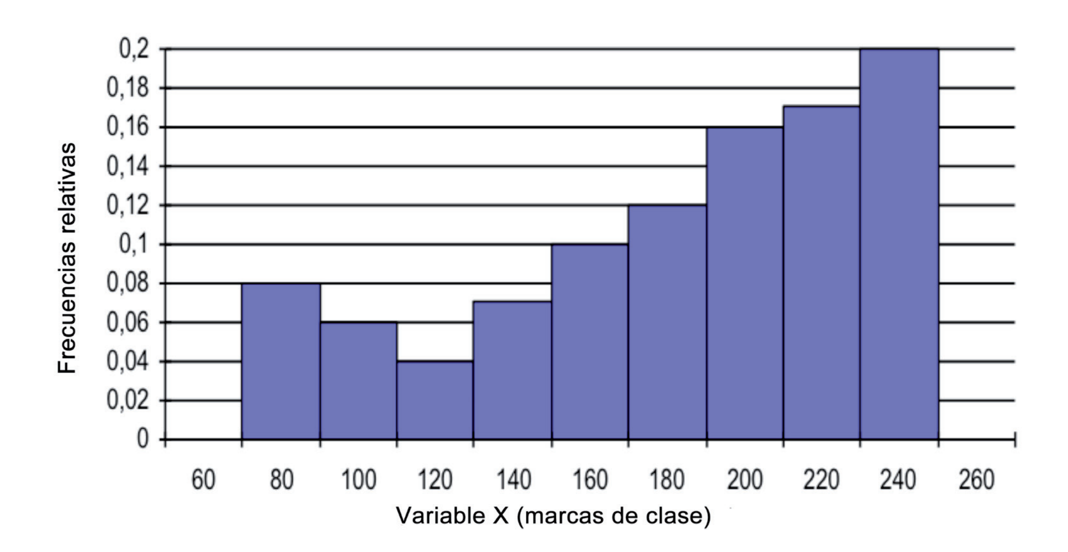
Se pide:

a) ¿De qué tipo de variable es el objeto de estudio?
b) Mostrar en forma de tabla de frecuencias el conjunto de los datos recogidos.
c) ¿Qué porcentaje de familias se gastan más de 150.000 euros?
d) El 65 % de familias que menos se gasta, ¿qué cantidad de dinero como máximo desembolsa?

1*b*) Saber clasificar las variables estadísticas.
1*c*) Saber analizar y realizar tablas de frecuencias de un conjunto de datos.
1*e*) Saber interpretar y construir los principales gráficos estadísticos.
1*f*) Conocer los conceptos y saber realizar los cálculos de las medidas de tendencia central y de dispersión. Concretar con la aplicación del coeficiente de variación de Pearson en aquellas situaciones que lo requieran.

Ejercicio 6

En el siguiente histograma se representa la distribución del dinero que durante el último mes se han gastado los trabajadores de una empresa en dietas:



a) Determina, sabiendo que hay 200 trabajadores.
b) La tabla de frecuencias que muestra los datos que tenemos.
c) La cantidad media que se han gastado, la más frecuente y la cantidad que tenían como máximo, el 50 % de los trabajadores que menos cobraban.
d) Calcula e interpreta el rango de la distribución así como el rango intercuartílico.
e) Calcula el mínimo del 20 % de los empleados con mayor cantidad de dietas. ¿Qué porcentaje del total de la empresa corresponde a este grupo?
f) El intervalo centrado en la cantidad media en que se encuentran el 75 % de los datos. ¿Es, pues, el sueldo medio muy representativo del conjunto de las dietas?
g) En el mes siguiente, la empresa decidió aumentar las dietas de todos los trabajadores un 5 %. Además, les dio una prima de 50 euros en concepto de productividad. Calcula el salario medio, el salario más frecuente y el salario que tenían como máximo, el 50 % de los trabajadores que menos cobran el mes siguiente.

h) De las dietas de otra empresa, que pertenece al mismo sector, se sabe que la media aritmética de sus trabajadores es de 120 euros, con una varianza de 2,5 euros. ¿Qué empresa tiene una dieta media más representativa? Razona la respuesta.

1*b*) Saber clasificar las variables estadísticas.
1*g*) Conocer los principales estadísticos que miden la forma de los datos a partir de los gráficos.

Ejercicio 7

Se quiere lanzar al mercado un nuevo producto cerámico y la empresa que lo crea estudia el tiempo de publicidad, en segundos, que otras empresas han utilizado para promocionar un producto similar. A continuación se puede ver para cada empresa la duración y los anuncios realizados:

Empresa 1

Duración	Número de anuncios
0-20	3
20-25	17
25-30	13
30-40	9
40-60	8

Empresa 2

Duración	Número de anuncios
0-20	1
20-25	5
25-30	13
30-40	5
40-60	2

Empresa 3

Duración	Número de anuncios
0-20	4
20-25	6
25-30	7
30-40	5
40-60	3

Empresa 4

Duración	Número de anuncios
0-20	3
20-25	17
25-30	13
30-40	9
40-60	8

Para realizar el estudio, calcula:

a) La duración media de cada empresa.
b) ¿Tienen todas las distribuciones la misma forma? Comenta el resultado.

1*h*) Saber calcular e interpretar el índice de Gini, así como saber realizar la curva de Lorenz para medir la equidad de un reparto.

Ejercicio 8

Dos compañías de venta de coches tienen maneras diferentes de pagar a sus trabajadores. La compañía A lo hace mediante un sueldo fijo mensual y la compañía B mediante un porcentaje sobre las ventas efectuadas. La distribución de los salarios por categorías es la siguiente:

COMPAÑIA A	
Sueldo (centenares de euro)	Número de trabajadores
26	10
39	10
52	40
247	20
260	10
273	10

COMPAÑIA B	
Sueldo (centenares de euro)	Número de trabajadores
4	10
5	10
6	40
7	20
26	10
27	10

a) Basándose únicamente en las observaciones, ¿en qué compañía el sueldo medio fluctúa menos o tiene los repartos más equitativos? Justifica el resultado mediante el análisis estadístico del reparto.

b) ¿En cuál de las dos compañías el sueldo es más homogéneo o concentrado? Se debe obtener el resultado también de forma gráfica.

1*a*) Conocer los conceptos básicos de las variables estadísticas.
1*e*) Saber interpretar y construir los principales gráficos estadísticos.
1*f*) Conocer los conceptos y saber realizar los cálculos de las medidas de tendencia central y de dispersión. Concretar con la aplicación del coeficiente de variación de Pearson en aquellas situaciones que lo requieran.

Ejercicio 9

La distribución de edades del Censo Electoral de Residentes a 1 de enero de 1999 para las comunidades autónomas de Aragón y Canarias, en tantos por ciento, es la siguiente:

Edades	Aragón	Canarias
16–18	3,55	4,35
18–30	21,56	29,99
30–50	31,63	35,21
50–70	28,14	21,97
70–90	15,12	8,48

a) Representa sobre los mismos ejes de coordenadas los datos de la distribución de la edad para las dos comunidades autónomas (emplea distinto trazo o distintos colores). ¿Qué conclusiones obtienes a la vista del gráfico?
b) Calcula la edad media para las dos comunidades. Compáralas. ¿Qué indican estos resultados?
c) ¿En qué comunidad las observaciones son más dispersas?
d) Si los datos de edades fueron: Aragón: 10, 10, 10, 10, 20, 30, 40, 30, 40, 50, 60, 40, 40, 40, 60, 70, 80, 70, 80, 90 , 70, 50, 40, 90. Canarias: 20, 30, 40, 40, 140, 50, 40, 30, 40, 30, 50, 60, 40, 30, 30, 40, 30, 40, 30, 40 , 30, 50, 60, 70. Obten un gráfico que nos muestre la dispersión de los datos en el mismo gráfico.

Ayudas

En este apartado se presentarán las ayudas a emplear en caso de ser necesario a la hora de realizar los ejercicios y problemas. Es conveniente no hacer un abuso excesivo de estas ayudas, es decir, antes de emplear la ayuda hay que pensar el problema al menos durante unos 10-15 minutos. Después se consultará la ayuda de tipo 1 y se intentará resolver el ejercicio con esta ayuda. Si no es posible resolverlo, entonces se consultará la ayuda de tipo 2; y en último término la solución.

Ayudas Tipo 1

Exercicio 1

Lo que se necesita para resolver este ejercicio, es primeramente conocer los tipos de variables que existen. A continuación puedes ver una clasificación de los tipo de variables.

Las *variables cualitativas* son aquellas que no se pueden medir, es decir, aquellas que toman valores a los que no se puede asignar ningún número. Expresan cualidades o categorías. Además pueden ser:

a) *Ordinales*: se pueden ordenar.
b) *Nominales*: no hay preferencias entre unas y otras.

Las *variables cuantitativas*, por el contrario, son medibles, es decir, los valores que se observan pueden expresarse de forma numérica. Estas variables pueden clasificarse en:

a) *Discretas*, cuando toman sus valores en un conjunto finito o numerable.
b) *Continuas*, cuando pueden tomar cualquier valor en un intervalo.

Exercicio 2

Lo que se necesita para resolver este ejercicio, es conocer primeramente los tipos de variables que existen para elegir la correcta y el tipo de gráfico correspondiente. La clasificación del tipo de variables, como ya se conoce del ejercicio anterior es:

– Las *variables cualitativas (Ordinales o Nominales)*.
– Las *variables cuantitativas (Discretas o Continuas)*.

Según el tipo de variable, el gráfico correspondiente será:

- Para las variables cualitativas: diagrama de barras o diagrama de sectores.
- Para las variables cuantitativas:

  Discretas: Diagrama de barras o sectores.
  Continuas: Histograma.

Los primeros pasos serán saber qué tipo de variable es, ya que este elemento afectará a la elección tanto del tipo de tabla de frecuencias como la elección del tipo de gráfico.

Queda claro que es una variable numérica. Por lo tanto, puede ser continua o discreta. En este caso, como los datos hacen referencia al número de hijos será cuantitativa discreta.

Con estas informaciones, se puede pasar a resolver el problema.

Ejercicio 3

Lo que se necesita para resolver este ejercicio, de la misma forma que el anterior, es conocer los tipos de variables que existen para elegir la correcta y el tipo de gráfico correspondiente. En este caso, los que aparecen son datos numéricos continuos. Por este motivo, lo que se trabaja es la creación de representaciones gráficas como son los histogramas en los dos apartados.

Por otra parte, hay que pensar cómo crear las clases para hacer este tipo de problemas y se puede hacer con el conocimiento de los siguientes elementos:

Se llama *marca de clase* a la media aritmética de los dos extremos del intervalo. Es evidentemente el valor central del intervalo ya que equidista de los extremos. Se denota por c_i. Se calcula $c_i = \frac{L_{i-1} + L_i}{2}$.

Se llama *amplitud de un intervalo* o recorrido a la distancia que hay entre los extremos.

Se llama *densidad de frecuencia absoluta* de un intervalo al cociente entre la frecuencia absoluta del intervalo y su amplitud.

Método de la raíz: Según este método el número de clases es igual a la raíz cuadrada del número de datos:

Número de clases = $\sqrt{número \;\; de \;\; datos}$.

El siguiente paso es calcular la amplitud de los intervalos.

Por lo que, la amplitud $\approx \frac{recorrido}{número\ de\ clases}$.

Con esta información, se puede empezar sin problemas la solución del problema.

Ejercicio 4

Hay que recordar, sin embargo, que los diferentes valores que puede tomar la variable estadística se denotan mediante x_i. En este caso, ordenándolos de menor a mayor, $x_1 = 5,\ x_2 = 6,\ x_3 = 7, x_4 = 10,\ x_5 = 11,\ x_5 = 12,\ x_6 = 15$.

Se llama *frecuencia absoluta del valor* x_i al número de veces que aparece repetida la observación en la recopilación de datos. Se representa por n_i. La frecuencia absoluta del valor x_2 es 2 ($n_2 = 2$), pues el dato 6 se repite dos veces en el conjunto de los datos de la muestra.

Se llama *frecuencia relativa del valor* x_i al cociente entre la frecuencia absoluta de x_i y el número total de datos *n*. Se representa por f_i y, evidentemente, es la proporción en que se encuentra el valor x_i dentro del conjunto de datos en tanto por uno; $f_i = \frac{n_i}{n}$. En el ejemplo $f_2 = \frac{n_2}{n} = \frac{2}{20} = 0{,}1$. Por tanto, el 10 % de los datos son seises.

Es importante remarcar que para calcular frecuencias acumuladas, a las que llamaremos F_i como frecuencia relativa acumulada y N_i como frecuencia absoluta acumulada, es necesario que las variables a estudiar sean *ordenables*, es decir, debe ser posible establecer una relación de orden entre las variables. Sin embargo, no tiene ningún sentido realizar dichos cálculos.

Estas definiciones permiten resumir los datos. Sin embargo, la manera más adecuada para sintetizar los datos es mediante lo que se denomina tabla de frecuencias. En ella aparecen distribuidas los datos según las frecuencias. Al mismo tiempo refleja todos los conceptos mencionados con anterioridad.

Ejercicio 5

a) En los ejercicios anteriores ya hemos visto que es necesario conocer la clasificación de las variables.

b) La clasificación del tipo de variables, como ya se conoce del ejercicio anterior es:

– Las *variables cualitativas* (*Ordinales o Nominales*).
– Las *variables cuantitativas* (*Discretas o Continuas*).

c) Para completar la tabla de frecuencias debemos conocer:

- Saber crear la tabla de datos continuos. En este caso, los intervalos ya los tenemos, solo tenemos que añadir la marca de clase.
- Completar la tabla con las diversas frecuencias n, f, N y F.
- Además, hay que conocer los pasos para crear la tabla si no conocemos los intervalos, pero esto es un problema que no tenemos en este caso.

d) Debemos buscar en la tabla el valor en los intervalos. En este caso, el intervalo que tiene el máximo en 150.000.

e) Se pide el percentil 65 Los percentiles dividen la distribución en 100 partes (99 divisiones). P1,..., P99, correspondientes al 1 %,..., 99 %. En este caso, el valor correspondiente al percentil 30, tiene un 30 % de los datos superiores o iguales a él.

Ejercicio 6

Como primera ayuda recordar que:

- Hay que saber el tipo de variable. En este caso es una variable cuantitativa continua, ya que se muestra un gráfico con formato de histograma.
- Además, recuerda que en el eje de las ordenadas, lo que aparece es la frecuencia relativa, ni la absoluta ni ninguna de las acumuladas. Esta suposición se basa en que el gráfico no está en todo momento aumentando.
- El formato de la tabla de frecuencias tendrá la forma:

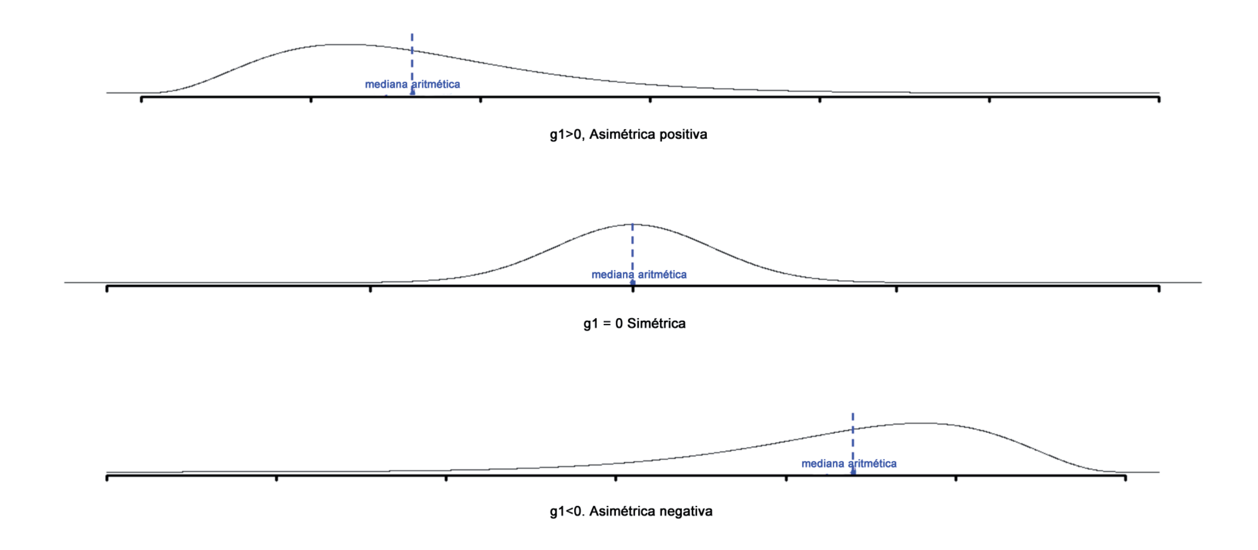
$[L_{i-1}, L_i)$	ni	Ni	fi	Fi

Ejercicio 7

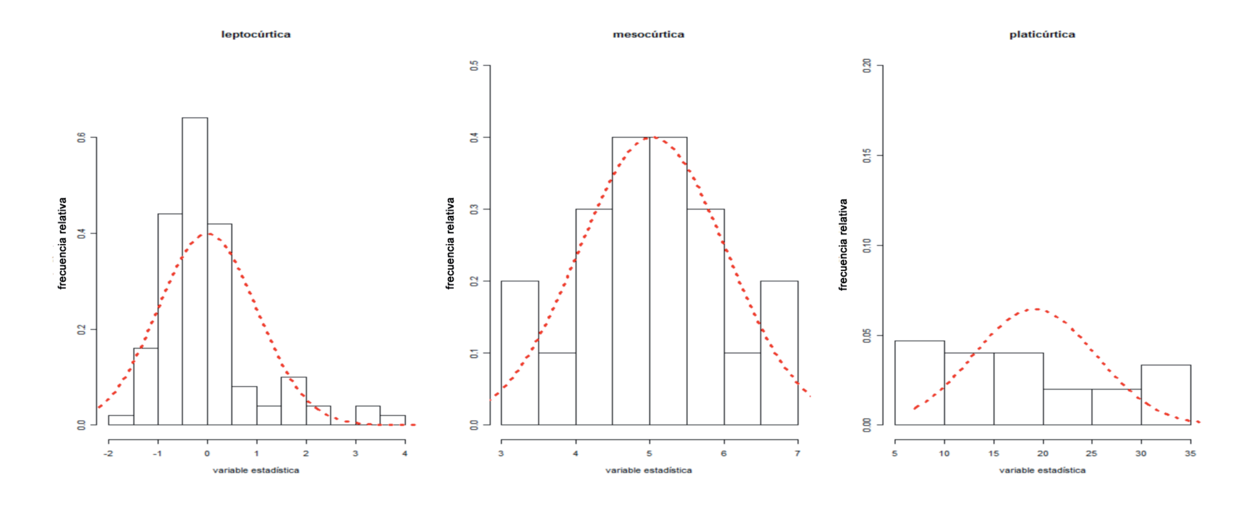
a) Por lo que respecta al cálculo de la media aritmética, hay que tener en cuenta que es una variable continua y que hay que utilizar la marca de clase en cada caso.

b) Respecto a la forma de las distribuciones, tenemos que trabajar los coeficientes de asimetría y curtosis, además de crear los gráficos para ver la forma de las distribuciones. En este caso se puede utilizar el diagrama de barras aplicado a cada empresa para ver la desviación respecto a la distribución normal.

La forma de los gráficos puede ser respecto a la simetría:



Respecto al apuntamiento:



Ejercicio 8

a) Respecto a lo que se nos pregunta en el primer apartado, debemos calcular el coeficiente de variación de cada una de las compañías y luego realizar la comparación.

b) En el segundo apartado se pregunta el índice de concentración de Gini. Para calcularlo, se seguirá la siguiente información: se supone que se tiene una distribución de rentas $(x_i \cdot n_i)$ donde i toma los valores de 1 hasta k (por ejemplo x_i *son los sueldos* y n_i el número de personas que cobran ese sueldo) de la que se formará una tabla con las columnas siguientes:

1) Los productos, $x_i \cdot n_i$ indicarán la renta total percibida por los n_i rentistas de renta individual x_i.

2) Las frecuencias absolutas acumuladas N_i.

3) Los totales acumulados u_i que se calculan de la siguiente forma:

$$u_1 = x_1 n_1$$
$$u_2 = x_1 n_1 + x_2 n_2$$
$$u_3 = x_1 n_1 + x_2 n_2 + x_3 n_3$$
$$u_4 = x_1 n_1 + x_2 n_2 + x_3 n_3 + x_4 n_4$$

....

$$u_k = x_1 n_1 + x_2 n_2 + x_3 n_3 + x_4 n_4 + \ldots\ldots\ldots\ldots + x_k n_k$$

Por tanto, se puede decir:

$$u_j = \sum_{i=1}^{j} x_i \cdot n_i \text{ para cualquier valor de } j \text{ desde 1 hasta } k.$$

4) La columna total de frecuencias acumuladas relativas, que se expresa en tanto por ciento y que se representa por p_i, vendrá dada por la siguiente notación:

$$p_i = \frac{N_i}{n}$$

5) La columna de renta acumulada relativa, que se expresa en tanto por ciento y que se representa por la expresión:

$$q_i = \frac{u_i}{u_k}$$

Por lo tanto, se puede hacer la tabla:

x_i	n_i	$x_i n_i$	N_i	u_i	$p_i = \frac{N_i}{n}$	$q_i = \frac{u_i}{u_k}$	pi - qi
x_1	n_1	$x_1 n_1$	N_1	u_1	p_1	q_1	$p_1 - q_1$
x_2	n_2	$x_2 n_2$	N_2	u_2	p_2	q_2	$p_2 - q_2$
...	...	...	...	...	...	...	...
x_k	n_k	$x_k n_k$	N_k	u_k	100	100	0

Ejercicio 9

a) Los datos en este caso son muy importantes, ya que se puede ver que se muestran los datos agrupados pero en diferente amplitud. Por este motivo, se ha de representar la densidad de los datos, no directamente los datos que se nos presentan.

b) Para obtener los datos de la media aritmética, hay que tener en cuenta el mismo elemento que se ha comentado con anterioridad, que son datos agrupados. Si la variable está agrupada en intervalos el concepto no cambia. En este caso, se asignan las frecuencias a las marcas de clase y se procede de la misma manera que en el caso de no agrupados.

c) El estudio de la dispersión está relacionada con el cálculo de la desviación típica en el caso del trabajo de variables por separado, pero en este caso, para compararlas, se utiliza el coeficiente de variación de Pearson.

d) Una posibilidad es obtener el gráfico de cajas y bigotes.

Ayudas Tipo 2

En este apartado se presentarán las ayudas para emplear en caso de ser necesario a la hora de realizar los ejercicios y problemas, y tras consultar la ayuda de tipo 1.

Ejercicio 1

Aunque se conozca la clasificación de las variables, y se tenga suficiente información para clasificar los distintos apartados, se pueden añadir ejemplos de cada caso para compararlos con los que se piden:

– *Variables cualitativas nominales*: el sexo o el color.
– *Variables cualitativas ordinales*: estar bien, regular o enfermo y también estar lleno, medio lleno o vacío.
– *Variables cuantitativas discretas*: número de trabajadores en una empresa o número de edificios en una calle.
– *Variables cuantitativas continuas*: la altura de las personas, las calificaciones numéricas de un examen o la medida en centímetros de la fabricación de tablas.

Ejercicio 2

Ya conoces qué tipo de gráficos se debe utilizar en cada caso. Ahora tienes que seguir los siguientes pasos para hacer los gráficos:

– Crear la tabla de frecuencias correspondiente, que en este caso, como es una variable cuantitativa discreta, no será necesario crear intervalos y luego crear los gráficos correspondientes con sus frecuencias.

Primeramente, crearemos la tabla de frecuencias para poder crear los gráficos correspondientes:

xi	ni	Ni	fi	Fi
0	15	15	0,15	0,15
1	21	36	0,21	0,36
2	21	57	0,21	0,57
3	27	84	0,27	0,84
4	16	100	0,16	1,00
Total	100		1	

a) Respecto a las representaciones gráficas, ya que se refiere a datos discretos, debemos utilizar un gráfico que puede ser el de sectores o de barras. Se representa en el eje de abscisas las clases, que en este caso es el número de hijos, y en el eje de ordenadas la frecuencia correspondiente, que puede ser tanto la absoluta como la relativa (acumulada o no).

b) Para construir el polígono de frecuencias con las frecuencias acumuladas se utilizarán también los datos de la tabla de frecuencias y podrán ser tanto la N como la F.

Ejercicio 3

El siguiente paso, con los datos agrupados en intervalos, será crear la tabla de frecuencias agrupada como queda a continuación:

$[L_{i-1}, L_i)$	**ni**	**Ni**	**fi**	**Fi**
[3,25–3,75)	3	3	0,075	0,075
[3,75–4,25)	8	11	0,2	0,275
[4,25–4,75)	14	25	0,35	0,625
[4,75–5,25)	6	31	0,15	0,775
[5,25–5,75)	4	35	0,1	0,875
[5,75–6,25)	5	40	0,125	1
	N = 40			

En cada apartado debemos:

a) Crear un histograma de forma general.
b) Se creará un histograma con cuatro clases, sin realizar la separación general de los datos agrupados que, como ya se conoce, es la raíz de los datos.

Ejercicio 4

El formato de la tabla será:

x_i	n_i	N_i	f_i	F_i
Total				

Y una forma de saber si los datos se mantienen a lo largo del tiempo o varían sería el crear un gráfico de barras, por ser una variable numérica discreta.

Ejercicio 5

a) La variable de estudio es la cantidad de euros que se gastan las familias para comprar la segunda vivienda en euros. Esta información nos ayudará en la solución de los próximos apartados.

b) El formato de la tabla será:

Euros	Marca	Familias (n_i)	f_i	N_i	F_i

c) Se puede utilizar la frecuencia relativa acumulada y restar a 1 el valor del F_i anterior.

d) Se pide el percentil 65.

En distribuciones agrupadas es necesario determinar el intervalo $[L_{i-1}, L_i)$ en el que se encuentra el cuantil. Este intervalo se determina siguiendo exactamente los mismos procedimientos mencionados en el apartado anterior; se realiza el mismo que en el caso de datos no agrupados. La diferencia radica en que se obtendrá un intervalo en lugar de un valor.

Una vez se tiene el intervalo $[L_{i-1}, L_i)$, el cuartil se calcula:

$$Cuartil = L_{i-1} + \frac{\frac{a}{100} \cdot n - N_{i-1}}{n_i} a_i \text{ donde,}$$

L_{i-1} Límite inferior de la clase del percentil
N_{i-1} Es la frecuencia absoluta acumulada de la clase «anterior» a la clase del percentil
n_i Es la frecuencia de la clase del percentil
a_i Es la amplitud de la clase del percentil

Ejercicio 6

Todos los apartados siguientes al de la creación de la tabla dependen de esta, que ayuda a calcular cada uno de los estadísticos.

a) Nos preguntan: la media, la moda y la mediana.
b) El rango, como ya debes saber es la diferencia entre máximo y mínimo valor de la variable, y el rango intercuartílico es la diferencia entre el cuartil primero y tercero.
c) Nos preguntan el percentil 80, ya que nos habla de los valores más altos y a partir de este, se calcula el porcentaje.
d) Se debe aplicar el teorema de Thebyshev a partir de la desviación típica.
e) Aplicación de las propiedades de la media, la moda y la mediana, donde todos los factores que suman, restan, multiplican o dividen a la variable les afectan. En este caso: se multiplicaría por 0.05, por el 5 % y se sumaría a los tres estadísticos 50 euros.
f) Lo que se pide es comparar la variabilidad o la dispersión de dos muestras diferentes. En estos casos, lo más correcto es calcular el coeficiente de variación de Pearson, CV. Por este motivo, es necesario calcular tanto la media como la desviación típica de las variables.

Ejercicio 7

Para el apartado *b*), es necesario conocer la forma de los coeficientes de asimetría y curtosis.

De esta manera se obtiene el *coeficiente de asimetría de Fisher*.

$$g_1 = \frac{m}{s^3} = \frac{\dfrac{\sum_{i=1}^{k}(x_i - \overline{X})^3 n_i}{n}}{\left(\sqrt{\dfrac{\sum_{i=1}^{k}(x_i - \overline{X})^2 n_i}{n}}\right)^3}$$

Hay que notar que como la desviación típica es positiva, el signo del coeficiente de Fisher será el mismo que el de *m*. Y por lo tanto:

si $g_1 = 0$	*la distribución es simétrica*
si $g_1 > 0$	*la distribución es asimétrica positiva*
si $g_1 < 0$	*la distribución es asimétrica negativa*

Así pues, cuando $g_1 < 0$, se dice que la distribución presenta asimetría a la izquierda (o negativa) y entonces, de las dos ramas de la curva que separa la ordenada que pasa por la media aritmética, la de la izquierda es más larga que la de la derecha. Lo opuesto ocurre si $g_1 > 0$.

Del mismo modo que en el caso del estudio de la asimetría, hay un coeficiente que permite clasificar los datos según la curtosis. En este caso, el coeficiente no es tan intuitivo, por lo que únicamente se dará la definición y su interpretación. Como en el caso de la otra medida de forma, este indicador tampoco tiene dimensión.

$$g_2 = \frac{\dfrac{\sum_{i=1}^{k}(x_i - \overline{X})^4 n_i}{n}}{\left(\dfrac{\sum_{i=1}^{k}(x_i - \overline{X})^2 n_i}{n}\right)^2} - 3$$

La idea del apuntamiento de una distribución de datos sale de la comparación de la frecuencia de los valores centrales de una distribución con la frecuencia de los

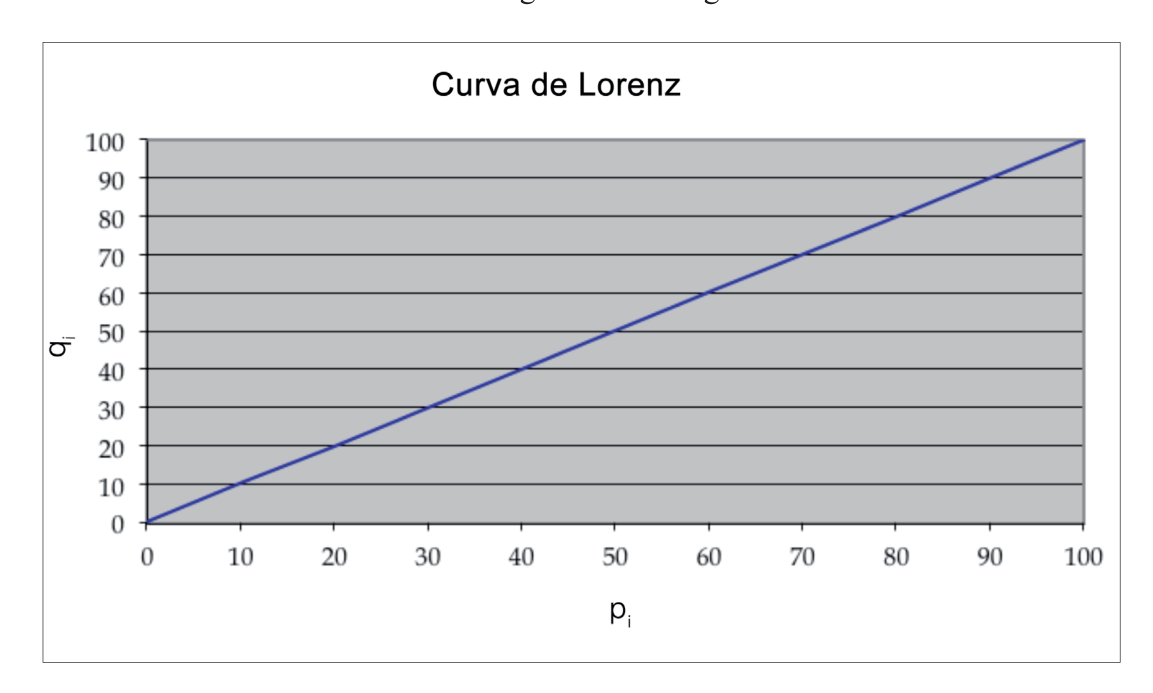
valores centrales en un modelo teórico normal que tenga la misma media y la misma desviación típica que la distribución que se está estudiando.

Como en un modelo normal se cumple que $\frac{\frac{\sum_{i=1}^{k}(x_i - \overline{X})^4 n_i}{n}}{s^4} = 3$, entonces, una distribución será:

mesocúrtica (normal)	si $g_2 = 0$
leptocúrtica	si $g_2 > 0$
platicúrtica	si $g_2 < 0$

Ejercicio 8

Como ayuda final, se puede comentar el tipo de gráfico que se ha de presentar. Es la curva de Lorenz. La forma de este gráfico es la siguiente:



Se deben representar los valores de q_i frente a los valores de p_i.

La línea central es la línea de equidad de los datos, que nos marcará el nivel de concentración.

Ejercicio 9

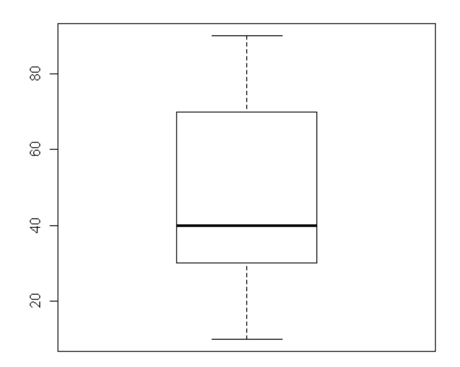
a) La obtención de las densidades será el cociente, en cada caso, de la frecuencia absoluta del intervalo entre su amplitud.

b) Un diagrama de cajas y bigote (conocido también como *Box and whisker plot* en inglés), es una representación gráfica de los datos que permite determinar con mucha facilidad y de una manera visual la tendencia central, la variabilidad, la asimetría y la existencia de valores anómalos de un conjunto de observaciones. De alguna manera, se puede decir que es uno de los gráficos que más y mejor resumen los conjuntos de datos.

El diagrama de cajas emplea el resumen de los 5 números: la menor observación, la mayor observación, el primer cuartil, la mediana y el tercer cuartil. Estos 5 números permiten construir la versión más simple del *Box plot*, el cual está formado por:

*Una caja (*box*) central* que representa las observaciones comprendidas entre el primer y el tercer cuartil. Los dos extremos de la caja son los cuartiles, y una línea interior y vertical que parte la caja en dos partes, corresponde a la mediana. Es obvio, pues, que la caja comprende el 50 % de las observaciones.

*Bigotes (*whiskers*)*: El gráfico se completa en esta versión del *Box plot*, con dos líneas a ambos lados de la caja que unen el primer cuartil con la menor observación, y el tercer cuartil con la observación mayor.



Soluciones

Ejercicio 1

Clasifica las siguientes variables, justificando el por qué de la elección:

a) Color de los coches.
b) Marcas de ordenadores.
c) Longitud de carreteras en metros.
d) Nivel de estudios.
e) Número de hijos de una familia.
f) Número de alumnos de estadística en una carrera.
g) Metros de altitud de las montañas.
h) Profesiones de las personas.
i) Sueldo mensual de los trabajadores de las empresas del sector cerámico.

Solución

a) Es una variable cualitativa nominal: color A, color B, color C, etc.
b) Es una variable cualitativa nominal: marca X, marca Y, marca Z, etc.
c) Es una variable cuantitativa continua: 1.93, 1.935, 1.76, 1.67, etc.
d) Es una variable cualitativa ordinal: sin estudios, elementales, etc.
e) Es una variable cuantitativa discreta: 0, 1, 2, 3, etc.
f) Es una variable cuantitativa discreta: 0, 1, 12, 3033, 5004, etc.
g) Es una variable cuantitativa continua: 36.1, 36.51, 36.512, 36.78, 37.1, 39.12, etc.
h) Es una variable cualitativa nominal: médico, profesor, payaso, etc.
i) Es una variable cuantitativa continua: 1200.50, 1165.43, 1500.23, etc.

Ejercicio 2

Actualmente, se está estudiando en las distintas comunidades autónomas el número de hijos por familia para estudiar la natalidad. Uno de los trabajadores que está haciendo las encuestas, recoge los datos de su barrio donde hay 100 familias. Ha obtenido los siguientes datos:

1	3	3	0	4	3	1	4	0	0
2	1	0	3	1	2	1	4	1	2
3	3	4	2	0	4	3	0	2	3
1	3	4	2	2	4	4	4	2	1
4	2	1	1	0	1	1	2	3	0

3	3	3	1	1	3	3	0	2	3
4	3	0	3	1	2	2	1	2	3
3	2	1	3	1	3	4	4	4	1
3	0	3	1	0	4	3	2	3	2
1	2	0	2	0	0	2	2	3	4

a) Construye el gráfico que consideres más adecuado con las frecuencias acumuladas.
b) Construye el polígono de frecuencias con las frecuencias acumuladas.

Solución

El primer paso será saber qué tipo de variable es, ya que este elemento afectará a la elección tanto del tipo de tabla de frecuencias como del tipo de gráfico.

Queda claro que es una variable numérica. Por lo tanto, puede ser continua o discreta. En este caso, ya que los datos hacen referencia a número de hijos será cuantitativa discreta.

Con estas informaciones, se puede pasar a resolver el problema.

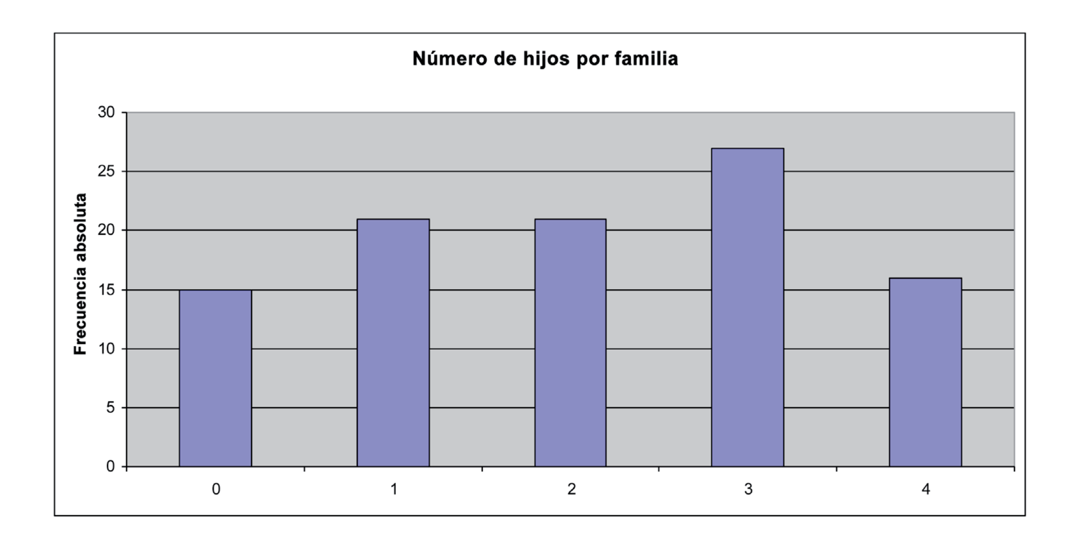
Primeramente, crearemos la tabla de frecuencias para poder crear los gráficos correspondientes:

xi	ni	Ni	fi	Fi
0	15	15	0,15	0,15
1	21	36	0,21	0,36
2	21	57	0,21	0,57
3	27	84	0,27	0,84
4	16	100	0,16	1,00
Total	100		1	

a) Respecto a las representaciones gráficas, como se refiere a datos discretos, debemos utilizar un gráfico que puede ser el de sectores o de barras. En ningún caso utilizaremos el histograma, ya que se usará para los datos continuos.

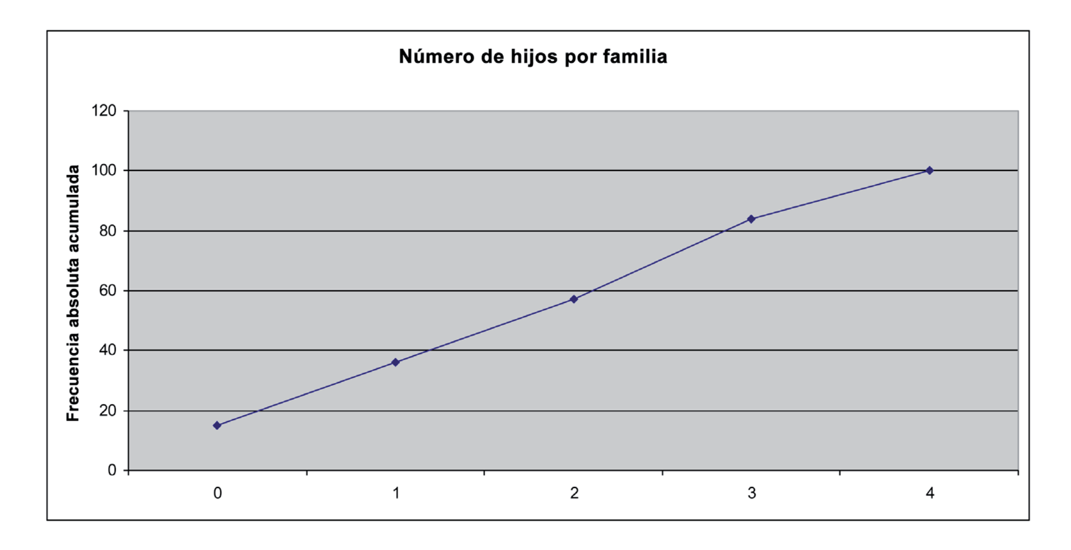
Se representa en el eje de abscisas las clases que, en este caso, es el número de hijos, y en el eje de ordenadas la frecuencia correspondiente, que puede ser tanto la absoluta como la relativa (acumulada o no).

El resultado de representar la frecuencia absoluta en un diagrama de barras es el siguiente:



b) Para construir el polígono de frecuencias con las frecuencias acumuladas, se utilizarán también los datos de la tabla de frecuencias y podrán ser tanto la N como la F.

Como se puede ver a continuación, lo que se utiliza como resolución es la frecuencia absoluta acumulada en el eje de ordenadas.



Ejercicio 3

Los sueldos, en miles de euros mensuales, de 40 empresarios del sector de la construcción del año 2007 son:

3,9	4,7	3,7	5,6	4,3	4,9	5,0	6,1	5,1	4,5
5,3	3,9	4,3	5,0	6,0	4,7	5,1	4,2	4,4	5,8
3,3	4,3	4,1	5,8	4,4	4,8	6,1	4,3	5,3	4,5
4,0	5,4	3,9	4,7	3,3	4,5	4,7	4,2	4,5	4,8

Se quiere estudiar si realmente son bastante altos y cuál es su distribución. Para conseguirlo:

a) Representa gráficamente la información recogida.
b) Crea la misma representación en 4 clases para poder diferenciar de forma más clara los tipos de sueldos.

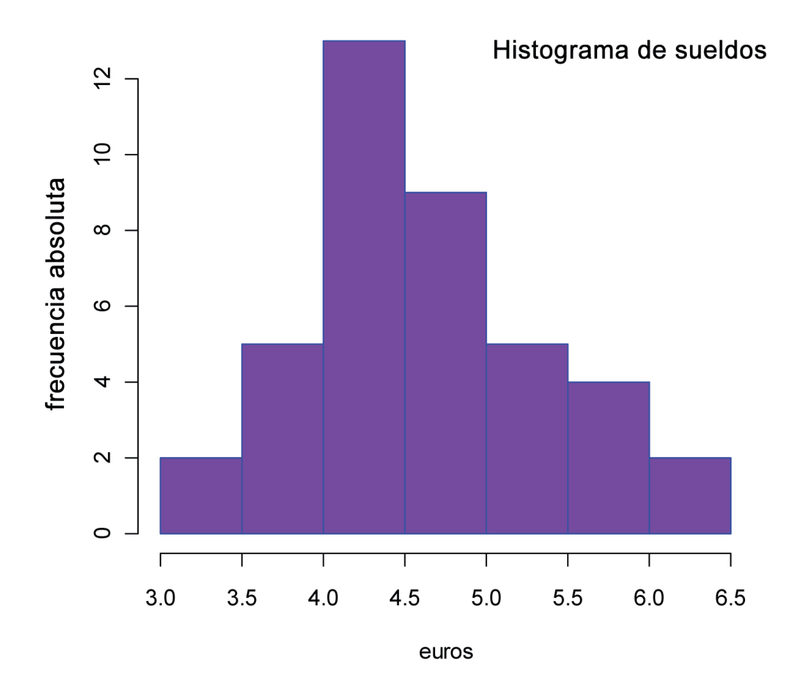
Solución

El primer paso será saber qué tipo de variable es, ya que este elemento afectará a la elección tanto del tipo de tabla de frecuencias como del tipo de gráfico. Queda claro que es una variable numérica continua.

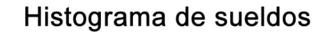
El siguiente paso es agrupar los datos en intervalos y crear la tabla de frecuencias agrupada:

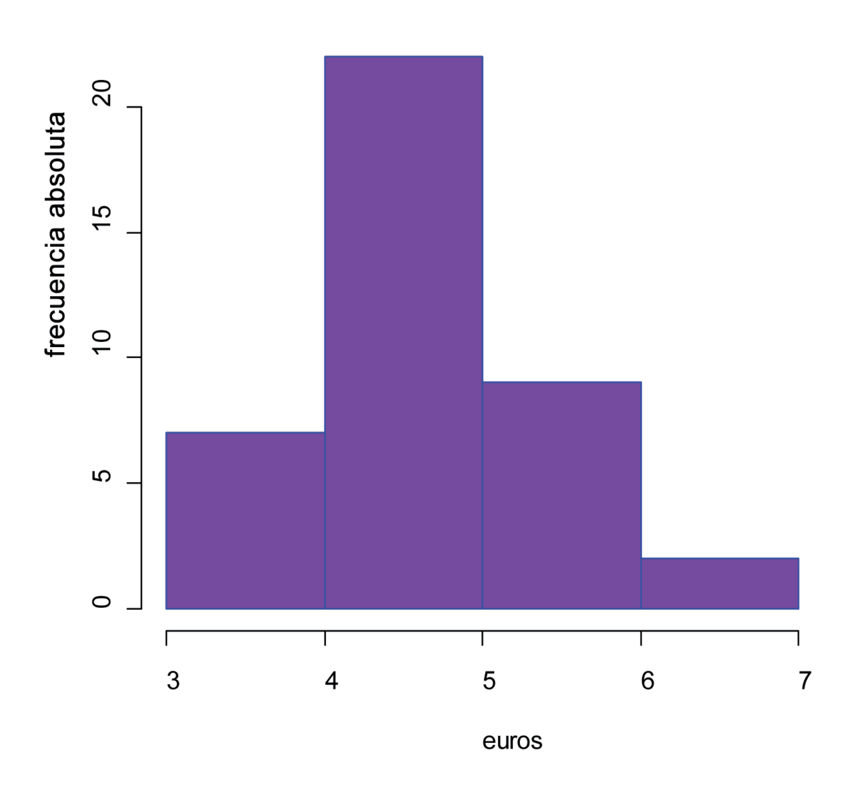
$[L_{i-1}, L_i)$	ni	Ni	fi	Fi
[3,25–3,75)	3	3	0,075	0,075
[3,75–4,25)	8	11	0,2	0,275
[4,25–4,75)	14	25	0,35	0,625
[4,75–5,25)	6	31	0,15	0,775
[5,25–5,75)	4	35	0,1	0,875
[5,75–6,25)	5	40	0,125	1
	N = 40			

a) Un posible resultado, será el siguiente histograma:



b) Como se pidn cuatro clases, el histograma pasará a ser el siguiente:





Ejercicio 4

La recopilación de 20 datos correspondientes al número de llamadas de teléfono registradas en una empresa durante los días de preparación de material para una feria de muestras durante el período de 9 a 12 horas.

15,5, 10, 5, 5, 6, 5, 6, 5, 6, 7, 10, 10, 12, 11, 11, 12, 15, 12, 15

Se quiere estudiar si realmente hay variación a lo largo de los días de las llamadas que se reciben. Por este motivo se pide confeccionar una tabla de frecuencias que recoja esta información.

Solución

Hay que recordar, sin embargo, que los diferentes valores que puede tomar la variable estadística se denotan mediante x_i. En este caso, ordenándolos de menor a mayor, $x_1 = 5$, $x_2 = 6$, $x_3 = 7, x_4 = 10$, $x_5 = 11$, $x_5 = 12$, $x_6 = 15$.

Se llama *frecuencia absoluta del valor* x_i al número de veces que aparece repetida la observación en la recopilación de datos. Se representa por n_i. La frecuencia absoluta del valor x_2 es 2 ($n_2 = 2$), pues el dato 6 se repite dos veces en el conjunto de los datos de la muestra.

Se llama *frecuencia relativa del valor* x_i al cociente entre su frecuencia absoluta x_i y el número total de datos *n*. Se representa por f_i y evidentemente, es la proporción en que se encuentra el valor x_i dentro del conjunto de datos en tanto por uno; $f_i = \frac{n_i}{n}$. En el ejemplo $f_2 = \frac{n_2}{n} = \frac{2}{20} = 0{,}1$. Por tanto, el 10 % de los datos son seises.

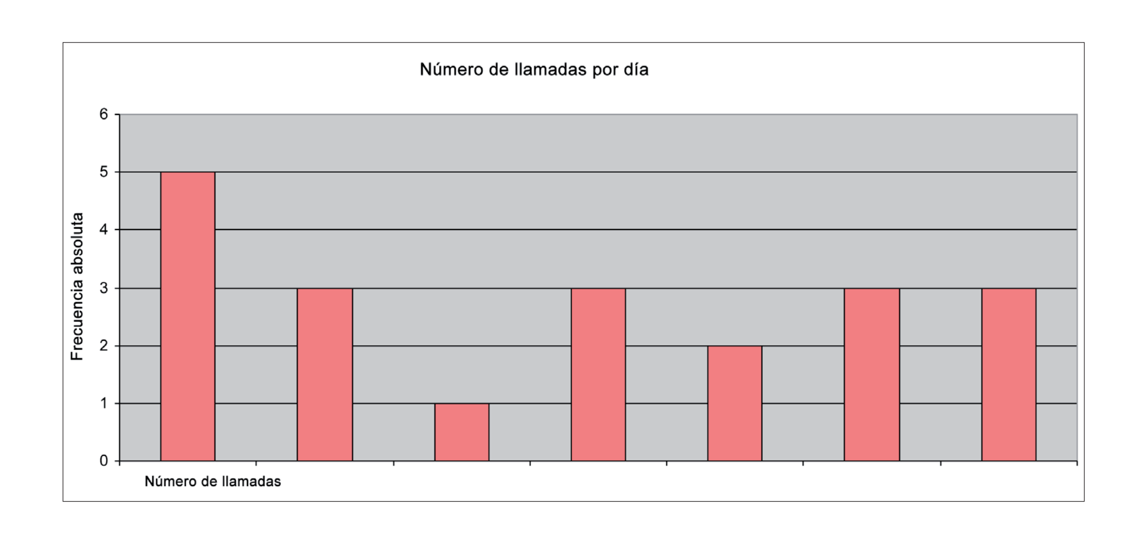
Es importante remarcar que para calcular frecuencias acumuladas, a las que llamaremos F_i como frecuencia relativa acumulada y N_i como frecuencia absoluta acumulada, es necesario que las variables a estudiar sean *ordenables*, es decir, debe ser posible establecer una relación de orden entre los valores de las variables. En otros casos, no tiene ningún sentido realizar dichos cálculos.

Estas definiciones permiten resumir los datos. Sin embargo, la manera más adecuada para sintetizar los datos es mediante lo que se denomina tabla de frecuencias. En ella aparecen distribuidas los datos según las frecuencias. Al mismo tiempo refleja todos los conceptos mencionados con anterioridad.

Con todos estos datos, el resultado de la tabla será el siguiente:

x_i	n_i	N_i	f_i	F_i
5	5	5	0,25	0,25
6	3	8	0,15	0,4
7	1	9	0,05	0,45
10	3	12	0,15	0,6
11	2	14	0,1	0,7
12	3	17	0,15	0,85
15	3	20	0,15	1
Total	20		1	

Si además, representamos la frecuencia absoluta, podemos ver que realmente no aumenta el número de llamadas, se mantiene bastante estable.



Ejercicio 5

Una empresa está haciendo el estudio del dinero que se gasta la gente para comprar una segunda casa como complemento de la primera vivienda. Se anotan los datos de los euros y el número de familias que han comprado este tipo de vivienda. A continuación se pueden ver los datos:

Euros	Familias
0-50000	2145
50000-75000	1520
75000-100000	840
100000-115000	955
115000-135000	1110
135000-140000	2342
140000-150000	610
150000-200000	328
>200000	150

Se pide:

a) ¿De qué tipo de variable es el objeto de estudio?
b) Mostrar en forma de tabla de frecuencias el conjunto de los datos recogidos.
c) ¿Qué porcentaje de familias se gasta más de 150.000 euros?
d) El 65 % de familias que menos se gasta, qué cantidad de dinero como máximo desembolsa?

Solución

a) La variable de estudio es la cantidad de euros que se gastan las familias para comprar la segunda vivienda en euros.

b) Para completar la tabla de frecuencias debemos conocer:

– El tipo de variable que se trabaja. En este caso es una variable cuantitativa continua.
– Saber crear la tabla de datos continuos. En este caso, los intervalos ya los tenemos, solo tenemos que añadir la marca de clase.
– Completar la tabla con las diversas frecuencias n, f, N y F.

La tabla que se nos pide será:

Euros	Marca	Familias (n_i)	f_i	N_i	F_i
0-50000	25000	2145	0,2145	2145	0,2145
50000-75000	62500	1520	0,152	3665	0,3665
75000-100000	87500	840	0,084	4505	0,4505
100000-115000	107500	955	0,0955	5460	0,546
115000-135000	125000	1110	0,111	6570	0,657
135000-140000	137500	2342	0,2342	8912	0,8912
140000-150000	145000	610	0,061	9522	0,9522
150000-200000	175000	328	0,0328	9850	0,985
>200000	200000	150	0,015	10000	1

c) Con la ayuda de la tabla, y con los datos de los intervalos, podemos ver cuáles son los casos superiores a 150.000. En este caso, por ejemplo, podemos utilizar la frecuencia relativa acumulada:

$$1 - 0.9522 = 0.0488, \text{ que será un } 4.88\ \%.$$

d) Se pide el percentil 65:

Los percentiles dividen la distribución en 100 partes (99 divisiones). P1,..., P99, correspondientes a 1 %,..., 99 %. En este caso, el valor correspondiente al percentil 30 tiene un 30 % de los datos inferiores o iguales a él.

En distribuciones agrupadas es necesario determinar el intervalo $[L_{i-1}, L_i)$ en el que se encuentra el cuantil. Este intervalo se determina siguiendo exactamente los mismos procedimientos mencionados en el apartado anterior; se realiza el mismo que en el caso de datos no agrupados. La diferencia radica en que se obtendrá un intervalo en lugar de un valor.

Una vez se tiene el intervalo $[L_{i-1}, L_i)$, el cuantil se calcula:

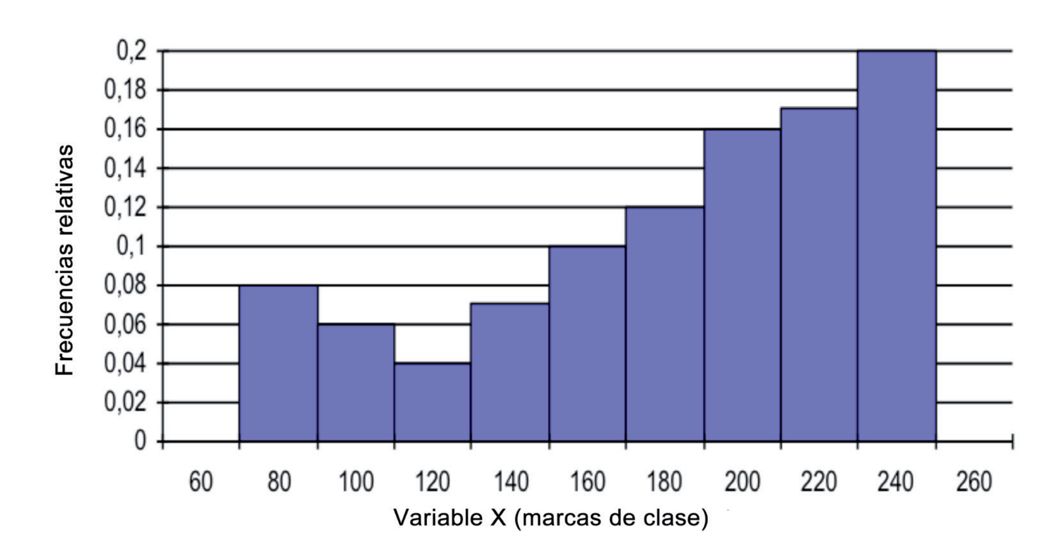
$$Cuantil = L_{i-1} + \frac{\frac{a}{100}\cdot n - N_{i-1}}{n_i} a_i \text{ donde,}$$

L_{i-1} Límite inferior
N_{i-1} Es la frecuencia absoluta acumulada de la clase «anterior» a la clase del cuartil
n_i Es la frecuencia de la clase del cuartil
a_i Es la amplitud de la clase del cuartil

En este caso, el valor del percentil 65 será de 133.738 euros.

Ejercicio 6

En el siguiente histograma se representa la distribución del dinero que durante el último mes se han gastado los trabajadores de una empresa en dietas:



a) Determina, sabiendo que hay 200 trabajadores.
b) La tabla de frecuencias que muestra los datos que tenemos.
c) La cantidad media que se han gastado, la más frecuente y la cantidad que tenían como máximo el 50 % de los trabajadores que menos cobraban.
d) Calcula e interpreta el rango de la distribución, así como el rango intercuartílico.
e) Calcula el mínimo del 20 % de los empleados con mayor cantidad de dietas. ¿Qué porcentaje del total de la empresa corresponde a este grupo?
f) El intervalo centrado en la cantidad media en que se encuentra el 75 % de los datos. ¿Es, pues, el sueldo medio muy representativo del conjunto de las dietas?

g) En el mes siguiente, la empresa decidió aumentar las dietas de todos los trabajadores un 5 %. Además, les dio una prima de 50 euros en concepto de productividad. Calcula el salario medio, el salario más frecuente y el salario que tenían como máximo el 50 % de los trabajadores que menos cobran.

h) De las dietas de otra empresa, que pertenece al mismo sector, se sabe que la media aritmética de sus trabajadores es de 120 euros, con una varianza de 2.5 euros. ¿Qué empresa tiene una dieta media más representativa? Razona la respuesta.

Solución

a) Tabla de frecuencias de datos agrupados a partir de un gráfico:

$[L_{i-1}, L_i)$	ni	Ni	fi	Fi
[70-90)	16	16	0,08	0,08
[90-110)	12	28	0,06	0,14
[110-130)	8	36	0,04	0,18
[130-150)	14	50	0,07	0,25
[150-170)	20	70	0,10	0,35
[170-190)	24	94	0,12	0,47
[190-210)	32	126	0,16	0,63
[210-230)	34	160	0,17	0,80
[230-250)	40	200	0,20	1
	N = 200		1	

b) $\overline{X}$ = 182 / Mo = 240 / Me = 193.750. Todos los estadísticos en euros.

c) Rango = 250 – 70 = 180. Para calcular el rango intercuartílico, hay que calcular primero el primer y tercer cuartil: C3 = 224,11 C1 = 150 Ri = 74,11

d) El sueldo mínimo es el P_{80} = 230. La proporción = 26,373 %.

e) El intervalo se encuentra aplicando el teorema de Thebyshev: [131.681, 232.319]; pues la desviación típica es de 50.319 euros.

f) $\overline{X}$ = 241 / Mo = 302 / Me = 253.43. Todos los estadísticos en euros.

g) Hay que calcular el coeficiente de variación de ambas observaciones. En la primera empresa, el coeficiente de variación es: CV = $\frac{S}{\overline{X}} = \frac{50,319}{182} = 0,276$ y en la segundo caso: CV = $\frac{\sqrt{2,5}}{120} = 0,013$.

Por tanto, la media aritmética de los sueldos de la segunda empresa es más representativo que la de la primera.

Ejercicio 7

Se quiere lanzar al mercado un nuevo producto cerámico y la empresa que lo crea estudia el tiempo de publicidad, en segundos, que otras empresas han utilizado para promocionar un producto similar. A continuación se puede ver para cada empresa la duración y los anuncios realizados:

Empresa 1

Duración	Número de anuncios
0-20	3
20-25	17
25-30	13
30-40	9
40-60	8

Empresa 2

Duración	Número de anuncios
0-20	1
20-25	5
25-30	13
30-40	5
40-60	2

Empresa 3

Duración	Número de anuncios
0-20	4
20-25	6
25-30	7
30-40	5
40-60	3

Empresa 4

Duración	Número de anuncios
0-20	3
20-25	17
25-30	13
30-40	9
40-60	8

Para realizar el estudio, calcular:

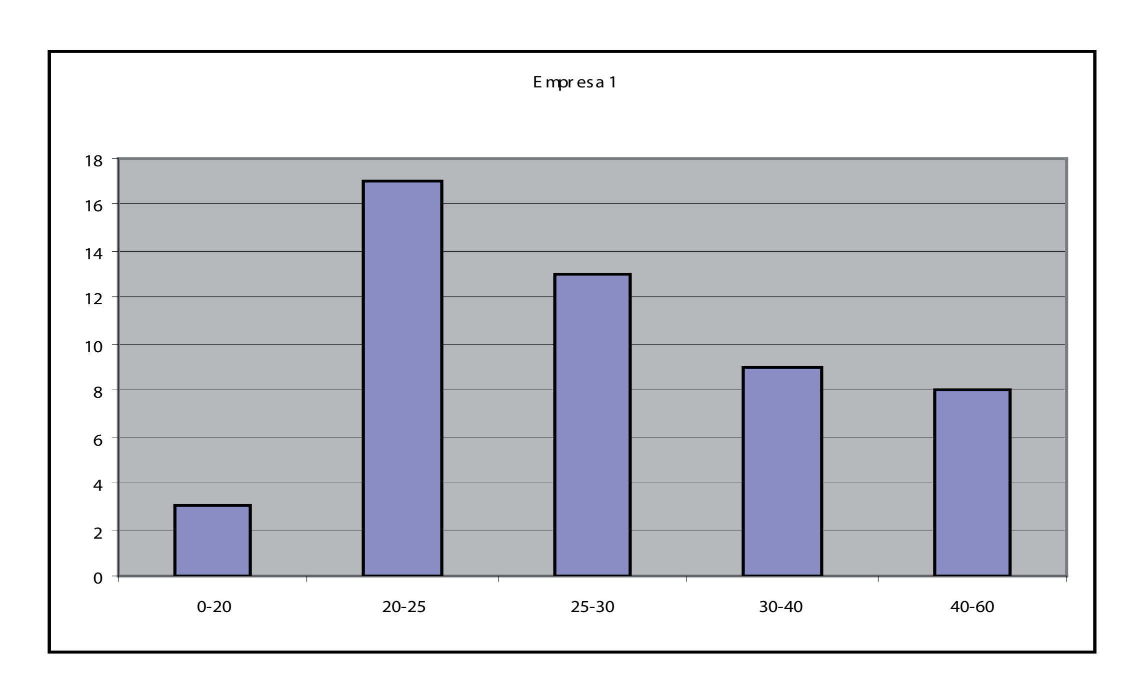
a) La duración media de cada empresa.
b) ¿Tienen todas las distribuciones la misma forma? Comenta el resultado.

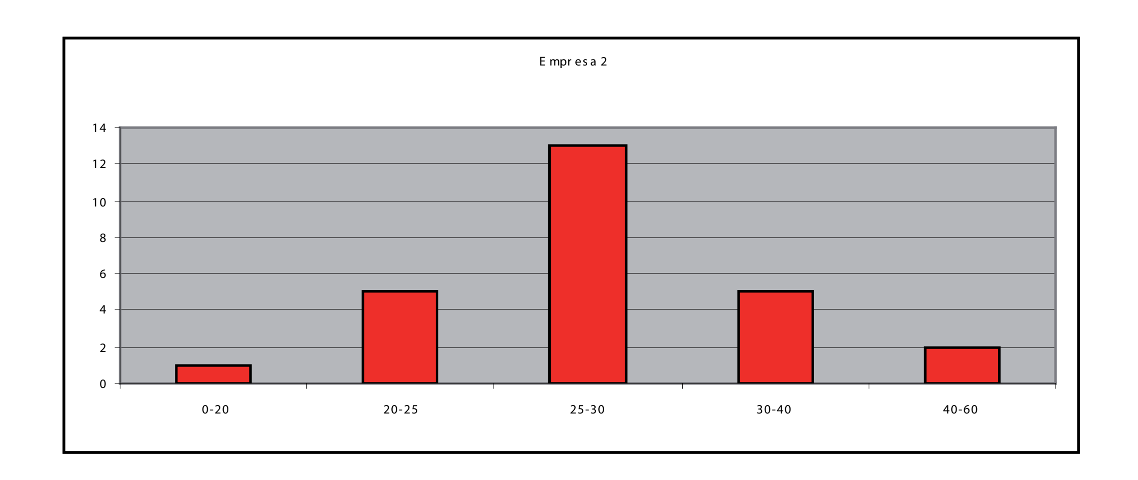
Solución

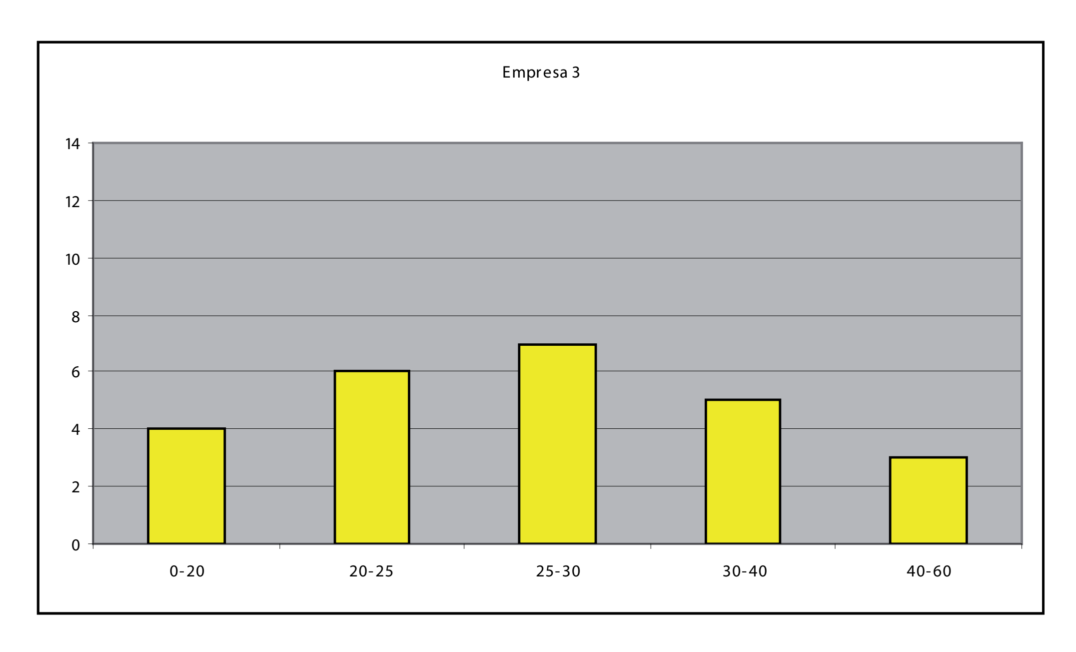
a) La media aritmética para cada caso será:

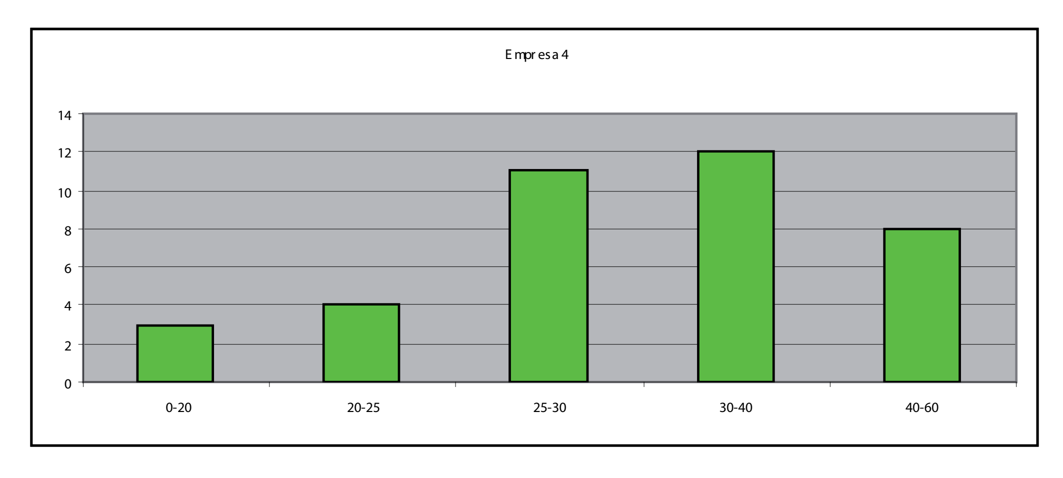
Empresa 1: 29.70 segundos.
Empresa 2: 29.04 segundos.
Empresa 3: 27.70 segundos.
Empresa 4: 32.70 segundos.

b) Representamos en forma de diagrama de barras o histograma de barras para ver la forma de la distribución:









Además, podemos calcular los valores de los coeficientes de asimetría y curtosis en cada caso, para ver claramente que:

– La empresa 1 es asimétrica a derechas.

– La empresa 2 es leptocúrtica.
– La empresa 3 es platicúrtica.
– La empresa 4 es asimétrica a izquierdas.

Asimetría (g1)	0,0506	1,4646	0,0000	-0,1231
Curtosis (g2)	-0,1875	2,4434	-1,2000	-2,7111

Ejercicio 8

Dos compañías de venta de coches tienen maneras diferentes de pagar a sus trabajadores. La compañía A lo hace mediante un sueldo fijo mensual y la compañía B mediante un porcentaje sobre las ventas efectuadas. La distribución de los salarios por categorías es la siguiente:

COMPAÑIA A	
Sueldo (centenares de euro)	Número de trabajadores
26	10
39	10
52	40
247	20
260	10
273	10

COMPAÑIA B	
Sueldo (centenares de euro)	Número de trabajadores
4	10
5	10
6	40
7	20
26	10
27	10

a) Basándose únicamente en las observaciones, ¿en qué compañía el sueldo medio fluctúa menos o tiene los repartos más equitativos? Justifica el resultado mediante el análisis estadístico del reparto.

b) ¿En cuál de las dos compañías el sueldo es más homogéneo o concentrado? Se debe obtener el resultado también de forma gráfica.

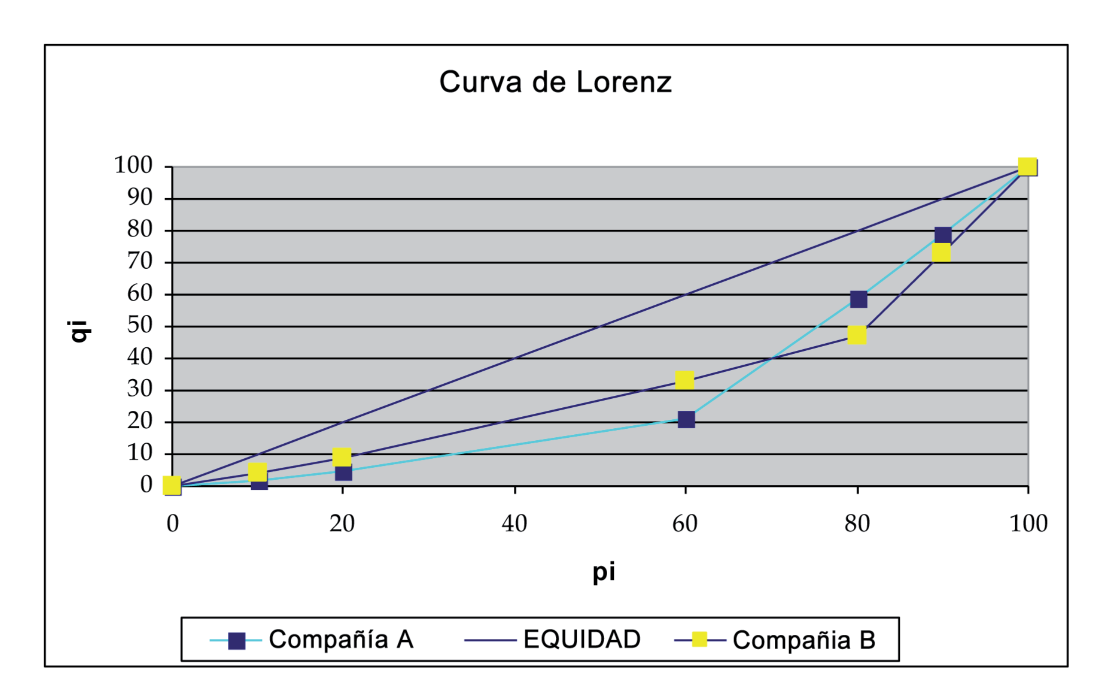
Solución

a) El sueldo medio de la compañía A es de 130 y el coeficiente de variación es de 83.2. El sueldo medio de la compañía B es de 10 y el coeficiente de variación

es de 6.88. Es decir, en la compañía A el sueldo medio es el menos representativo de los datos.

b) Las dos distribuciones de datos tienen el mismo índice de Gini: 0.361538. Por tanto, en las dos hay igual concentración.

A continuación se puede ver la representación de la curva de Lorenz para los dos casos:



Como se puede observar en el gráfico, las dos curvas de Lorenz se cruzan, por lo que, pese a tener distribuciones diferentes la concentración es la misma.

Ejercicio 9

La distribución de edades del Censo Electoral de Residentes a 1 de enero de 1999 para las comunidades autónomas de Aragón y Canarias, en tantos por ciento, es la siguiente:

Edades	Aragón	Canarias
16-18	3,55	4,35
18-30	21,56	29,99
30-50	31,63	35,21
50-70	28,14	21,97
70-90	15,12	8,48

a) Representa sobre los mismos ejes de coordenadas los datos de la distribución de la edad para las dos comunidades autónomas (emplea distinto trazo o distintos colores). ¿Qué conclusiones obtienes a la vista del gráfico?

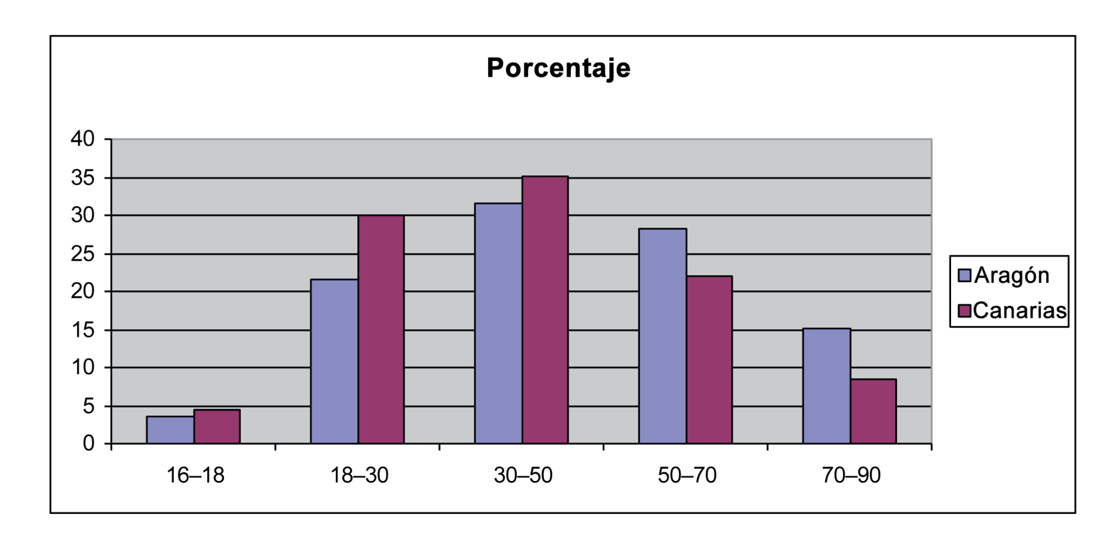
b) Calcula la edad media para las dos comunidades. Compáralas. ¿Qué indican estos resultados?

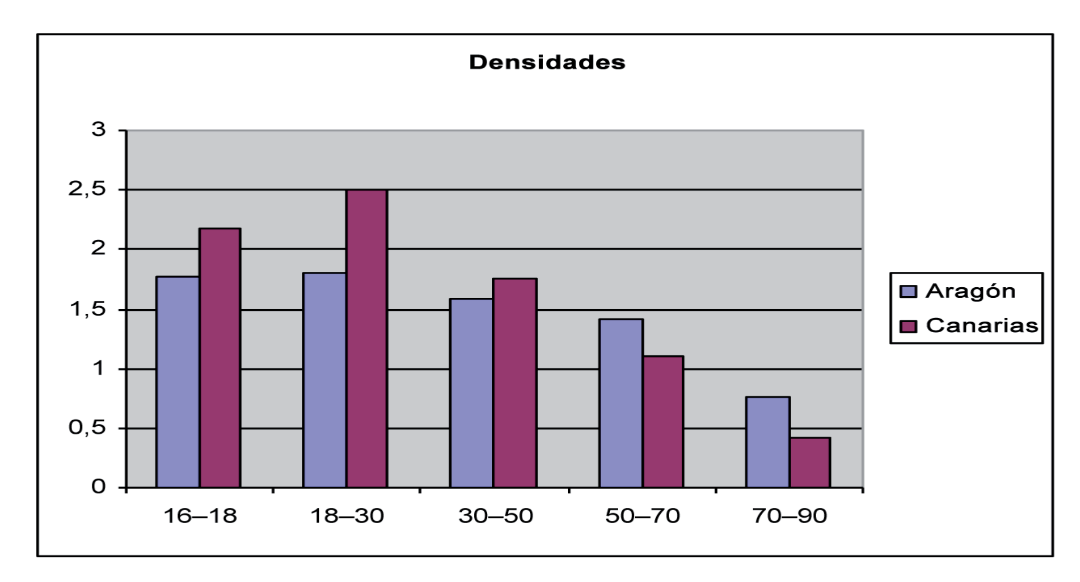
c) ¿En qué comunidad las observaciones son más dispersas?

d) Si los datos de edades fueron: Aragón: 10, 10, 10, 10, 20, 30, 40, 30, 40, 50, 60, 40, 40, 40, 60, 70, 80, 70, 80, 90 , 70, 50, 40, 90. Canarias: 20, 30, 40, 40, 140, 50, 40, 30, 40, 30, 50, 60, 40, 30, 30, 40, 30, 40, 30, 40 , 30, 50, 60, 70. Obten un gráfico que muestre la dispersión de los datos.

Solución

a) Se representan los dos conjuntos de datos teniendo en cuenta que los intervalos no tienen la misma amplitud y, por tanto, hay que calcular las densidades. Podemos ver la diferencia representándolos, tal y como aparecen los datos y la densidad, que será lo correcto:





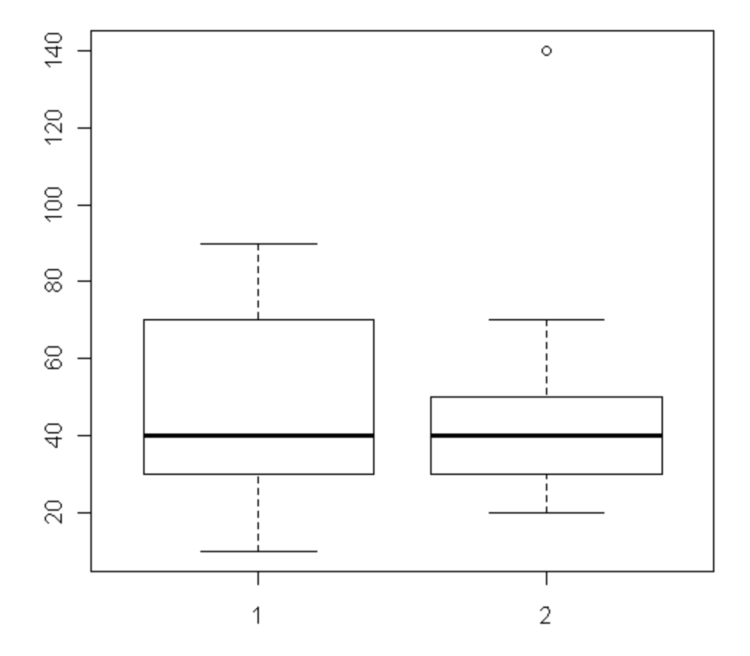
Claramente tenemos diferencias utilizando la densidad y aumenta la proporción en la comunidad de Aragón respecto a la de Canarias de gente de mayor edad.

b) La edad media de Aragón es de 47.41 años y la de Canarias de 41.98. Por los resultados, se podría decir que si las medias fueran representativas de los datos, la población de Aragón está un poco más envejecida que la de Canarias.

c) Para saberlo, hay que calcular los coeficientes de variación en ambos conjuntos de observaciones: CV Aragón = 40,75 %, CV Canarias = 42,56 %.

En consecuencia, en las islas Canarias las observaciones son más dispersas. No obstante, como los coeficientes de variación de ambas comunidades son muy altos, las medias aritméticas no serían muy representativas del conjunto de datos en ningún caso.

c) Lo que deberemos obtener es el gráfico de cajas y bigotes para las dos variables, siendo 1 = Aragón y 2 = Canarias:



La dispersión es major en Aragón que en Canarias.

UNIDAD 2

Estadística descriptiva bivariante

Introdución teórica

Normalmente, en cualquier investigación no se estudia una única variable de los individuos que forman la muestra (referencias bibliográficas 1, 5, 9, 12 y 16), sino que en muchas ocasiones son más. Así, si se desea estudiar el rendimiento de los trabajadores de una empresa, de cada trabajador puede ser útil conocer la edad, el sueldo, el nivel de estudios, las horas que trabaja, el número de personas que tiene a su cargo, etc. Es decir, para cada individuo de la muestra se obtiene un vector o registro en el que cada componente es el valor de una de las variables sujetas a estudio; en el ejemplo que se está considerando un vector asociado a un individuo sería:

(35 años, 24.500 €, diplomado, 47 horas semanales, 2 personas a su cargo...).

Este hecho origina que el investigador se plantee, además del estudio individualizado de cada una de las variables, el estudio conjunto de todas o de algunas ellas. De esta manera es posible conocer si existe algún tipo de relación funcional o estadística entre las variables. Así, las observaciones pueden manifestar que aquellas personas con más titulación tienen más personas a su cargo, o que a medida que va aumentando la edad de los trabajadores también lo hace el sueldo. Además, si esta relación existe puede que se pueda encontrar una «fórmula matemática» que relacione formalmente las variables.

Por otra parte, la nomenclatura cambia si se estudian conjuntamente diferentes variables. Así, si se realiza el estudio de dos variables se dice que se trabaja con variables bidimensionales, si son tres, variables tridimensionales, y si son más de tres, variables pluridimensionales.

Distribuciones estadísticas bidimensionales: tablas y gráficos

Cuando se desean estudiar dos características observables sobre una misma muestra o población, cada una de las variables que constituye la variable bidimensional (X, Y) se denomina componente o variable marginal de la misma, y puede ser tanto un atributo como una variable cuantitativa. En cualquier caso, al realizarse el trabajo de recogida de datos se obtiene un conjunto de pares ordenados del tipo:

$$\{(x_1, y_1), (x_1, y_1),..., (x_2, y_1), (x_2, y_1),..., (x_i, y_j),..., (x_i, y_j),..., (x_h, y_k),..., (x_h, y_k)\}$$

Por ejemplo, si se considerara X la variable días de estudio para un examen de Estadística y Y la nota obtenida para un conjunto de estudiantes, los datos recogidos serían del tipo:

$$\{(5,3) ; (6,5); (5,3) ; (6,5); (5,7)\}$$

En los datos, cada observación se repite un número de veces determinado. Así, una primera manera de representar el conjunto de datos es mediante la terna siguiente: $\{(x_i, y_i), n_{i,j}\}$ en la que:

- x_i representan los valores de la variable X
- y_i representan los valores de la variable Y
- $n_{i,j}$ es el número de veces que se repite cada dato (x_i, y_j), es decir, su frecuencia absoluta

Siguiendo con el ejemplo tenemos:

$x_1 = 5$	$y_1 = 3$	$n_{11} = 2$	
$x_1 = 5$	$y_3 = 7$	$n_{13} = 1$	
$x_2 = 6$	$y_2 = 5$	$n_{22} = 2$	La resot de $n_{i,j} = 0$

Por otra parte, es evidente que tener trescientos pares ordenados de observaciones aclara bien poco la información. No es posible observar casi nada. En consecuencia, es necesario representar los datos de manera que sean más comprensibles y facilitan el estudio. La manera de hacerlo es mediante tablas (tabla 1).

Tabla 1

X	Y	nij
x1	y1	n11
...	…	…
xi	yj	nij
.	.	.
. xh	yk	nhk

Para construir esta tabla ordenamos una de las variables, por ejemplo X, y vamos asociándole el valor correspondiente de la variable Y, así como su frecuencia absoluta conjunta. Si los datos fueran agrupados en intervalos, entonces la representación mediante esta tabla se realiza de forma similar. En ocasiones se utilizará la marca de la clase como representación del intervalo.

Esta tabla presenta y ordena los datos, sin embargo, en algunas ocasiones no es la tabla más adecuada y hay que construir la tabla de doble entrada o de contingencia.

Ejemplo 1

Valor del terreno	7	7	7	6,9	6,9	5,5	3,7	3,7	5,9	3,8	3,8	3,8	8,9	8,9	9,6	9,9	9,6	9,9	10	10	5,9	3,8	9,6	9,6	8,9	3,7	5,5	3,8	8,9	9,9
Coste de la vivienda	67	67	67	63	63	60	54	54	58	36	36	36	76	76	87	89	87	89	92	92	58	36	87	87	76	54	60	36	76	89

En 1999, los residentes de un pequeño pueblo estaban preocupados por el incremento del coste de la vivienda en la zona. El alcalde consideraba que los precios de la vivienda fluctuaban con los precios de los solares. Los costes de los terrenos y los de las viviendas (en miles de euros) sobre los que se construyeron las casas son los siguientes:

X	Y	$n_{i,j}$
3,7	54	3
3,8	36	4
5,5	60	2
5,9	58	2
6,9	63	2
7	67	2
7	67,15	1
8,9	76	4
9,6	87	4
9,9	89	3
10	92	2

Como se puede apreciar, los datos recogidos en la tabla anterior aportan poca información Se construyó, ya que no hay muchos pares diferentes, la tabla con las tres columnas. Se supondrá:

X = Valor del terreno
Y = Valor de la vivienda

Tabla de doble entrada o de contingencia

La tabla anterior, tal y como se ha comentado antes, algunas veces es incómoda y es preferible utilizar la tabla de doble entrada; la que permite extraer mucha más información de la distribución de datos. La tabla 2 presenta la forma de rectángulo, tal y como se puede observar a continuación:

Tabla 2

X \ Y	y1	y2	.....	yj	…..	yk	ni.
x1	n11	n12		n1j	…..	n1k	n1.
x2	n21	n22	…..	n2j	….	n2k	n2.
	..	.	.		.		.
xi	ni1	ni2		nij	….	nik	ni.
Xh	nh1	nh2		nhj	..	nhk	nh.
n. j	n. 1	n. 2		n. j	…	n. k	n

En la primera fila se sitúan las diferentes categorías o valores que toma una de las componentes, y en la primera columna los valores o las categorías relativas a la segunda (si es posible, ordenadas tanto la fila como la columna). De esta forma, cualquier número que aparece en una celda interior de la tabla de doble entrada es la frecuencia absoluta conjunta del dato bivariante, formado por los valores correspondientes ubicados en las correspondientes fila y columna. En algunas ocasiones también se suele representar en cada celda la frecuencia relativa conjunta, además de la absoluta.

Por otra parte, los valores que aparecen en la última columna y la última fila corresponden a las frecuencias absolutas de los valores de las variables de la primera columna y la primera fila respectivamente. Así, *ni.* representa la frecuencia absoluta del valor de la variable X, x_i.

Si los datos fueran agrupados en intervalos, entonces la representación mediante esta tabla se realizaría de forma similar, utilizando la marca de la clase como representación del intervalo.

Ejemplo 2

Con los mismos datos que en el ejemplo 1 anterior, la tabla de doble entrada queda:

X \ Y	36	54	58	60	63	67	67,15	76	87	89	92	ni.
3,7		3										3
3,8	4											4
5,5				2								2
5,9			2									2
6,9					2							2
7						2	1					3
8,9								4				4
9,6									4			4
9,9										3		3
10											2	2
n. j	4	3	2	2	2	2	1	4	4	3	2	29

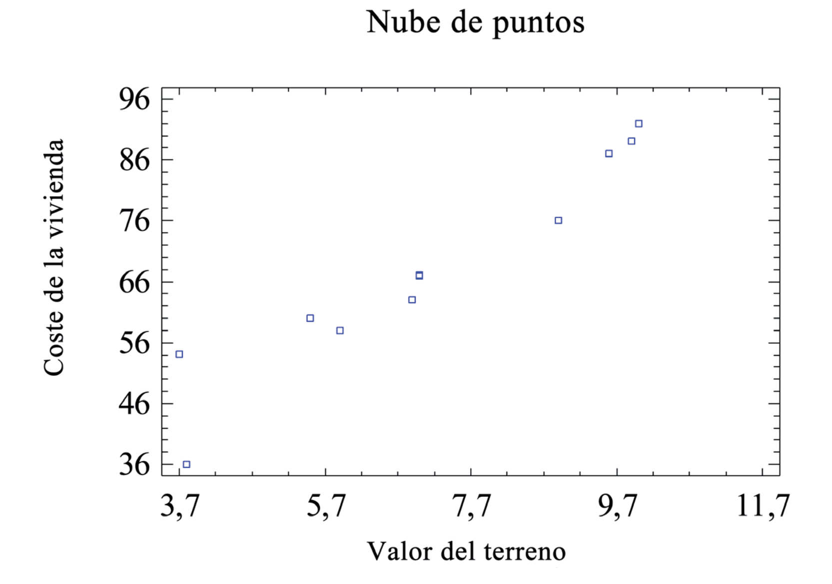
Las celdas vacías representan unas frecuencias absolutas conjuntas iguales a cero. Del mismo modo que ocurría con las distribuciones de datos unidimensionales, las representaciones gráficas facilitan la comprensión de la distribución con tan solo un vistazo.

Representaciones gráficas: diagrama de dispersión o nube de puntos

La representación gráfica de la distribución de frecuencias de una variable bidimensional (X, Y) varía sensiblemente según la naturaleza de las variables. Si las variables son discretas, la representación común de la distribución conjunta es la nube de puntos o diagrama de dispersión, el cual se construye situando sobre el eje horizontal de un sistema cartesiano los diferentes valores de la variable X, sobre la vertical los de la variable Y, y un punto en la posición (x_i, y_i) si es que esta observación tiene una frecuencia absoluta conjunta de 1. Si tuviera más de 1, hay diferentes posibilidades para representarlo: dibujar puntos de diferente superficies (la que representará la frecuencia), escribir la frecuencia junto al punto marcado, etc.

Ejemplo 3

Con los datos del ejemplo 1 que se está considerando relativo al valor del terreno y el costo de la vivienda, la nube de puntos es:



Distribuciones estadísticas marginales y condicionadas

Es evidente que la tabla de doble entrada mencionada en el epígrafe anterior ofrece mucha información. De hecho, es posible analizar cada variable componente de la variable conjunta, así como una variable condicionada a un valor concreto de la otra.

Distribuciones marginales

Si las variables X y Y son no agrupadas o cualitativas, la distribución marginal de X se obtiene de la tabla de doble entrada adjuntando a cada uno de los valores $x_1, x_2, ..., x_h$ de la variable estadística X, sus frecuencias absolutas, que vienen dadas en el ultima columna de la tabla. Asimismo, se obtiene la distribución marginal de Y. En este caso los valores de la variable $y_1, y_2, ..., y_k$ y sus frecuencias absolutas aparecen en la primera y la última fila respectivamente. Si las variables fueran agrupadas en intervalos, se realiza el mismo procedimiento tomando la marca de la clase como representante del intervalo, y por tanto, como valor de la variable estadística. Hay que decir que cada distribución marginal puede ser tratada estadísticamente como una variable unidimensional.

Ejemplo 4

En el ejemplo 1 se está considerando el valor del terreno y el coste de la vivienda:

Distribución marginal: Coste de la vivienda	
Y	n. j
36	4
54	3
58	2
60	2
63	2
67	2
67,15	1
76	4
87	4
89	3
92	2
	29

Distribución marginal: Valor del terreno	
X	ni.
3,7	3
3,8	4
5,5	2
5,9	2
6,9	2
7	3
8,9	4
9,6	4
9,9	3
10	2
	29

Distribuciones condicionadas

De la tabla de doble entrada, también es posible obtener, además de las distribuciones marginales, otras distribuciones. Si se asocia a los valores de Y las frecuencias correspondientes a la fila en la que está ubicado el valor $\mathbf{x_i}$ de **X**, resulta la distribución condicionada de Y a $\mathbf{x}_i$ (distribución de la variable $Y/_{X = x_i}$). Análogamente, pero teniendo presente las columnas en lugar de las filas, se obtendría la distribución de X condicionada a $\mathbf{y_j}$ de Y, (distribución de la variable $X/_{Y = y_j}$).

Ejemplo 5

Si en el ejemplo 1 se desea conocer la distribución del precio de la vivienda cuando el precio del solar es de 7000 euros, la distribución condicionada es $\text{Pr}\,ecio\ vivenda/_{\text{Pr}\,ecio\ solar\ =\ 7000}$ y se puede obtener la tabla de doble entrada:

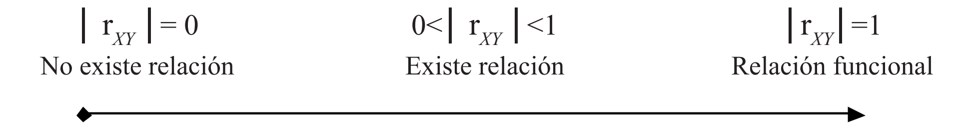
Distribución condicionada: Precio solar 7000 €	
$Y/_{X=7}$	$n_{.j}$
67	2
67,15	1
	3

Correlación lineal

Cuando se estudian dos variables estadísticas conjuntamente, es importante saber si hay algún tipo de relación entre ellas. Así, si se recogieran trescientos datos en que la primera variable fuera la altura de una persona y la segunda, el resultado de lanzar un dado, seguramente la intuición diría que las dos variables no tienen ningún tipo de relación entre sí. Si por el contrario, se consideraran las variables horas extra que trabaja una persona y el sueldo que cobra mensualmente, la relación cambiaría hasta el punto de conocer el sueldo de un individuo si se supiera las horas extras que hace. Se podría decir que las dos variables están ligadas por una relación funcional. Sin embargo, si se consideran las variables horas de preparación de un examen y nota obtenida, la intuición establecería que sí hay alguna relación entre ambas variables, siendo mucho más fuerte en el primer caso, pero más débil que en el segundo.

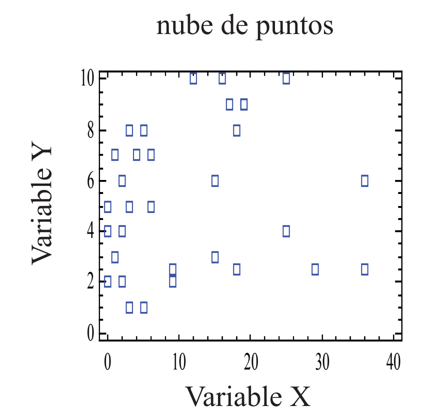
Como es evidente, las relaciones funcionales gozan de una fórmula que demuestra el tipo de relación. Por el contrario, para el resto de pares de variables no hay ninguna fórmula absoluta, a pesar de los lazos que existen en algunos casos. Para evidenciarlo surge el concepto de correlación, y el coeficiente de correlación r_{XY}.

Así, si dos variables tienen un relación muy fuerte, el valor absoluto de la correlación será muy próximo a 1 y en caso contrario será próximo al cero. Los casos 0 y 1 equivalen a no tener ningún tipo de relación y a tener una relación funcional. El vector siguiente lo resume:

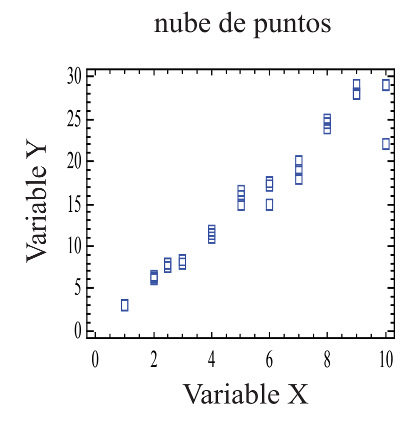


Cuando el tipo de relación funcional que se estudia entre las variables es una función lineal (una función del tipo y = ax + b), se habla de correlación lineal. A lo largo de la unidad, cuando se mencione el término correlación se considerará la correlación lineal, si no se explicita otra cosa.

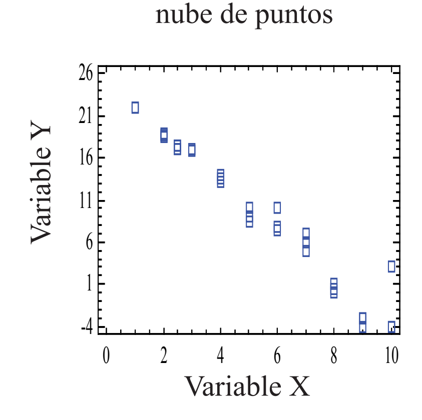
Una primera manera de observar la relación existente entre las variables X y Y son los gráficos de dispersión. Así, teniendo en cuenta lo expuesto al comienzo de este punto sobre la correlación:



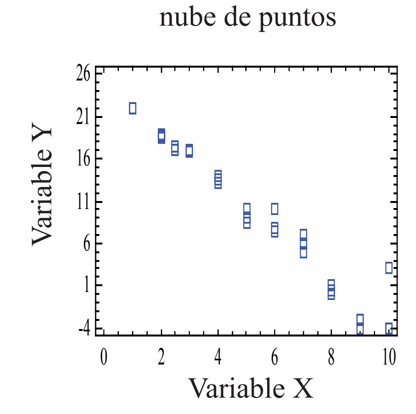
a) No existe correlación



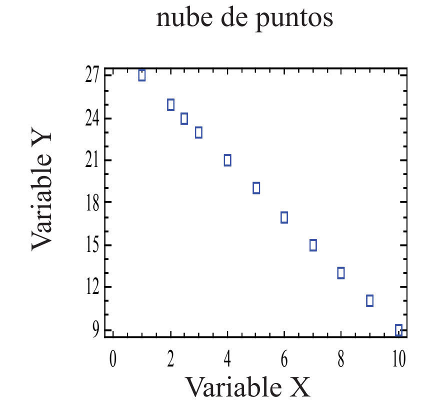
b) Correlación lineal positiva marcada



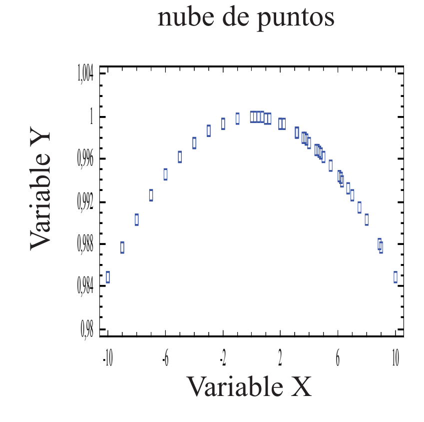
c) Correlación lineal positiva perfecta



d) Correlación lineal negativa marcada



e) Correlación lineal negativa perfecta



f) Correlación no lineal entre X i Y

Como se puede observar, en el ejemplo *e*) se detecta una relación entre las variables X y Y. Sin embargo, es evidente que no se trata de una relación lineal; pues estos tipos de relaciones determinan una nube de puntos similares a una línea recta. (Ejemplos *b*, *c*, *d* y *e*). En el ejemplo *a*), no se distingue ningún tipo de vínculo entre ambas variables; los puntos están muy dispersos.

Covarianza

El gráfico es una primera aproximación al estudio de la relación que existe entre las variables, pero únicamente aporta información de tipo intuitivo. El concepto que es necesario definir para poder decidir si hay o no relación lineal entre dos variables es el de *correlación lineal*. En primer lugar, debemos introducir el concepto de covarianza.

La covarianza es un estadístico (o un parámetro) por calcular, similar al de varianza y permite conocer si dos variables están relacionadas o no linealmente, se representa por S_{XY} y se calcula según la fórmula:

$$S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(x_i - \overline{X})(y_j - \overline{Y})\cdot n_{ij}}{n}$$

La interpretación de este estadístico es la siguiente:

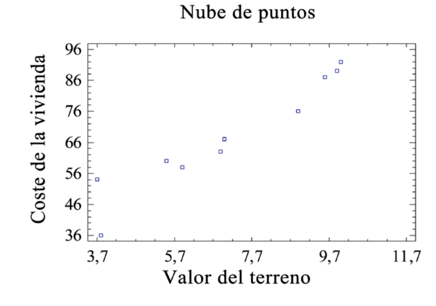
- Si $S_{XY} > 0$ ==> dependencia lineal directa (positiva), es decir a grandes valores de X corresponden grandes valores de Y (ejemplos *b*) y *c*).
- Si $S_{XY} > 0$ ==> incorrelacionadas, es decir no hay relación lineal (ejemplos *a*) y *f*).
- Si $S_{XY} > 0$ ==> dependencia lineal inversa o negativa, es decir a grandes valores de X corresponden pequeños valores de Y (ejemplos *d*) y *e*).

Ejemplo 6

En el caso que se está considerando el valor del terreno y el valor de la vivienda, la covarianza es:

Sabiendo que Valor terreno = $\overline{X}$ = 7,1586, Valor vivienda: = $\overline{Y}$ = 68.0052 y n = 29 se calcula:

$$S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(x_i - \overline{X})(y_j - \overline{Y})\cdot n_{ij}}{n} = \sum_{i=1}^{10}\sum_{j=1}^{11}\frac{(x_i - 7{,}1586)(y_j - 68{,}0052)\cdot n_{ij}}{29} = \mathit{42{,}1527\ €}$$



Así pues, el valor de la covarianza es coherente con la nube de puntos que ha obtenido: parece existir un relación directa o positiva entre valor del terreno y coste de la vivienda. Además, amayor valor del terreno, más coste de la vivienda.

Propiedades de la covarianza

- Si a todos los valores de la variable X, se les suma una constante *b* y a todos los valores de la variable Y una constante C, la covarianza no varía.

  Es decir, $S_{X+b\ Y+C} = S_{XY}$

- Si todos los valores de una variable *x* se multiplican por una constante *a* y todos los valores de la variable Y por una constante *b*, la covarianza queda multiplicada por el producto de las constantes.

  Es decir, $S_{a\cdot X\ b\cdot Y} = a \cdot b\ S_{XY}$

- A partir de las anteriores: si se tienen dos variables X y Y con covarianza S_{XY}, y dos transformaciones lineales de las variables de la forma X' = *ax* + *c*, i Y' = *by* + *d*, la nueva covarianza se relaciona con la anterior de la forma:

$$S_{a\cdot X\ b\cdot Y} = a \cdot b\ S_{XY}$$

Cálculo de la covarianza

Existe otra forma de obtener la covarianza mediante un cálculo más sencillo:

$$S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} - \overline{X} \cdot \overline{Y}$$

Se puede demostrar la equivalencia de ambas definiciones mediante procedimientos algebraicos elementales.

Ejemplo 7

Se verá a continuación un ejemplo de aplicación de esta última propiedad. De la siguiente tabla de doble entrada se determinará la covarianza:

X \ Y	1,6	1,7	1,8	
60	2	1	0	3
70	2	4	2	8
80	1	1	4	6
90	0	2	1	3
	5	8	7	20

En primer lugar se calculará la media aritmética de cada variable marginal:

$$\overline{X} = 74{,}5 \text{ y } \overline{Y} = 1{,}71$$

En segundo lugar, hay que calcular el primer sumando de S_{XY}, $\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n}$. Por ello, es necesario calcular primero los productos, sumarlos todos y luego dividir el resultado por el número total de datos:

60 * 1,6 * 2 =	192	Por tanto tendremos:
60 * 1,7 * 1 =	102	
70 * 1,6 * 2 =	224	$\sum_{i=1}^{h}\sum_{j=1}^{k} x_i \cdot y_j \cdot n_{ij} = 2554$
70 * 1,7 * 4 =	476	
70 * 1,8 * 2 =	252	$\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} = \frac{2554}{20}$
80 * 1,6 * 1 =	128	
80 * 1,7 * 1 =	136	
80 * 1,8 * 4 =	576	$S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} - \overline{X} \cdot \overline{Y} =$
90 * 1,7 * 2 =	306	
90 * 1,8 * 1 =	162	$= \frac{2554}{20} - 74{,}5 \cdot 1{,}71 =$
TOTAL SUMA =	2554	$= 127{,}7 - 127{,}395 = 0{,}305$

Correlación lineal

La covarianza permite discernir si dos variables X y Y tienen una relación positiva, negativa o cero, pero no aporta información del grado de dependencia de una variable respecto a la otra (referencias bibliográficas 6, 10 y 17). Además, la covarianza depende de las unidades de medida empleadas para X y Y –si, por ejemplo, X se mide en m^3 y Y en mm^3, cada desviación de X aumenta S_{XY} 10^9 veces–. Para hacer frente a estas dos dificultades se define el concepto ya introducido anteriormente de correlación lineal r_{XY}:

$$r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y}$$ siendo S_x y S_y las desviaciones típicas de X y Y.

Es evidente que, por definición, el coeficiente de correlación lineal informa de las mismas cosas que lo hace la covarianza. Además, cumple una propiedad muy importante, está acotado por 1 y por –1. Así pues r_{XY} , se caracteriza por:

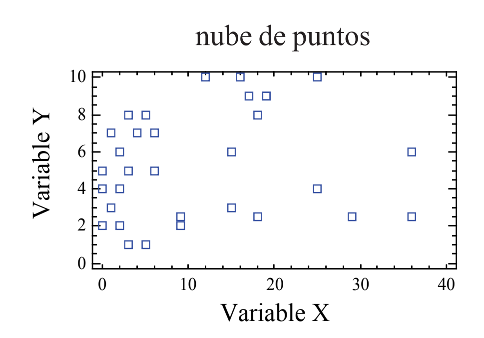
- Ser adimensional y siempre estar comprendido entre –1 y 1.
- Si hay relación lineal fuerte positiva, $r_{XY} > 0$ y está cerca de 1.
- Si hay relación lineal negativa fuerte, $r_{XY} < 0$ y está cerca de –1.
- Si no hay relación lineal r_{XY} será 0.

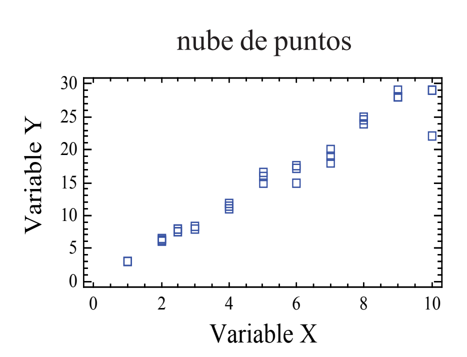
Ejemplo 8

En el ejemplo 1 que se está considerando, para calcular la correlación es necesario primero conocer las varianzas y la covarianza. Aprovechando los cálculos anteriores, se tiene: $S_{XY} = 42{,}1527$; $S_X = 18{,}1656$; $S_Y = 2{,}4242$. En consecuencia, el coeficiente de correlación lineal $r_{XY} = \dfrac{S_{XY}}{S_X \cdot S_Y} = \dfrac{42{,}1527}{18{,}1656 \cdot 2{,}4242} = 0{,}9572$. Por lo tanto, la relación lineal entre las dos variables es alta.

Recta de regresión. Calidad del ajuste

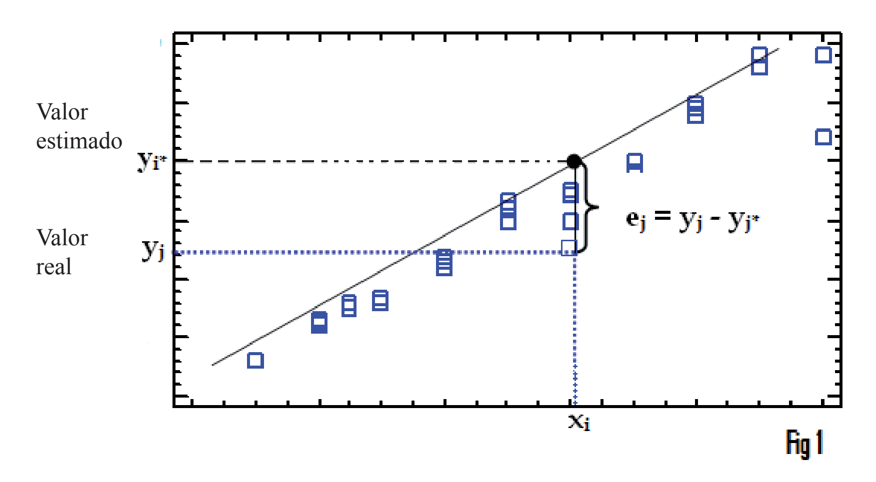
Como se ha expuesto anteriormente, cuando se estudian dos características simultáneamente sobre una muestra, se puede considerar que una de ellas influye sobre la otra de alguna manera. El objetivo principal de la regresión es descubrir el modo en que se relacionan. Un dibujo de la nube de puntos o diagrama de dispersión de la distribución puede indicar si es razonable pensar que puede haber una buena correlación lineal entre las dos variables.





En los diagramas anteriores se puede observar como en el de la derecha, una línea recta puede aproximarse a casi todos los puntos, mientras que en el otro, cualquier recta deja muchos puntos alejados de ella. Así, hacer un análisis de regresión lineal solo estaría justificado en el ejemplo de la derecha.

Como se puede ver en ambos diagramas, ninguna recta es capaz de pasar por todos los puntos. De todas las rectas posibles, la recta de regresión de Y sobre X es aquella que minimiza «el error de aproximación», considerando X como variable explicativa o independiente y Y como la explicada o dependiente. Pero ¿cómo se calcula la recta y se minimiza el error?



Se considera la recta $y = a + b$ X donde a y b donde a y b son parámetros. De este modo, la recta o función lineal es genérica (representa todas las funciones lineales posibles, únicamente hay que dar valores a los parámetros para obtener las infinitas rectas). Lo que se va a realizar consiste en encontrar los valores de los parámetros a y b, de modo que la recta se ajuste lo más posible a los puntos de la figura anterior. El método que se emplea para buscar los valores de los parámetros a y b es el de los mínimos cuadrados.

Usando técnicas de derivación se deduce que, de todos los posibles valores de a y de b, aquellos que minimizan la suma anterior son:

$$a = \bar{y} - \frac{s_{xy}}{s_x^2} \cdot \bar{x} \qquad \text{y} \qquad b = \frac{s_{xy}}{s_x^2}$$

NOTA: No hay que olvidar que si se conocen los datos también se conocen los términos: $\{\bar{y}, \bar{x}, S_{xy}, S_x^2\}$, y por tanto a y b serán números reales en el momento que se produzcan las sustituciones.

Así, sustituyendo en Y = $a + b$ X, la ecuación de la recta de regresión de Y sobre X es:

$$y = \left(\bar{y} - \frac{s_{xy}}{s_x^2} \cdot \bar{x} \right) + \left(\frac{s_{xy}}{s_x^2} \right) \cdot x$$

y también se puede escribir de la forma siguiente, recordando la ecuación de la recta punto-pendiente:

$$y - \bar{Y} = \frac{S_{XY}}{S_X^2} \cdot (x - \bar{X})$$

Si se hubiera tomado Y como variable independiente o explicativa, y X como dependiente o explicada, la recta de regresión que se necesita es la que minimiza errores de la X. Se llama recta de regresión de X sobre Y y se calcula fácilmente permutando los puestos de *x* e *y*, obteniéndose:

$$x - \overline{x} = \frac{s_{xy}}{s_y^2} \cdot \left(y - \overline{y} \right)$$

Ejemplo 9

En el ejemplo que se está considerando, se tiene que la variable independiente es el valor del terreno y el valor de la vivienda es la variable dependiente.

Por los estudios realizados a lo largo de la unidad se sabe que la relación es directa, pues la covarianza es positiva. Como la correlación obtenida ha sido un número cercano a 1, la relación lineal entre las dos variables es importante. Por tanto, el cálculo de la recta de regresión tiene sentido. Para calcularla, se utilizará la última expresión:

$$y - \overline{Y} = \frac{S_{XY}}{S_X^2} \cdot (x - \overline{X})$$. Así, $$\frac{S_{XY}}{S_X^2} = \frac{42{,}1527}{5{,}8768}$$ y la recta será:

$y - 7{,}1586 = \dfrac{42{,}1527}{5{,}8768}\ (x - 68{,}0052)$, y aislando la variable *y*, la recta $y = 16{,}6583 + 7{,}17274\,x$.

Por lo tanto:

Coste de la vivienda = 16,6583 + 7,17274 · Valor del terreno

Calidad del ajuste. Coeficiente de determinación

El *coeficiente de determinación lineal* se puede definir como el porcentaje de varianza de Y que se puede explicar por X y, se le suele llamar *Calidad o bondad del ajuste* porque valora la proximidad de la nube de puntos a la recta de regresión (o dicho de otro modo, como está de ajustada la nube de puntos a la recta de regresión).

En cuanto al cálculo del coeficiente de determinación, hay que definir previamente:

- La varianza de la variable Y que es explicada por la regresión lineal, llamada S_r^2, y que representa la variabilidad de la variable Y causada por las variaciones de la variable X.

- La varianza residual, que se representa por S_e^2, determina en qué medida difieren los valores ajustados por la recta de los valores observados. Es decir, se plantea medir la magnitud de los residuos.

Así:

$$S_r^2 = \sum_{i=1}^{h}\sum_{j=1}^{k}(y_j^* - \overline{y^*})^2 \frac{n_{ij}}{n} \quad \text{y} \quad S_e^2 = \sum_{i=1}^{h}\sum_{j=1}^{k}(y_j - y_j^*)^2 \frac{n_{ij}}{n}$$

Se puede demostrar matemáticamente que, en la regresión lineal de la variable Y sobre la variable X, la varianza de la variable Y se puede descomponer de la siguiente manera:

$$S_Y^2 = S_r^2 + S_e^2$$

Así pues, de la relación se deduce que cuanto mayor sea la varianza explicada por la regresión lineal (S_r^2) respecto de la varianza total, menor será la variabilidad del error de ajuste (S_e^2) y mejor será la bondad del ajuste.

Si ahora se divide la expresión anterior para S_Y^2, se obtiene: $1 = \frac{S_r^2}{S_Y^2} + \frac{S_e^2}{S_Y^2}$

Y retomando el *significado del coeficiente de correlación lineal (R^2)* como el porcentaje de varianza de Y que se puede explicar por X, se tiene:

$$R^2 = \frac{S_r^2}{S_Y^2} = 1 - \frac{S_e^2}{S_Y^2} \qquad \text{(En tanto por uno)}$$

De esta definición se pueden sacar algunas conclusiones:

- $0 \leq R^2 \leq 1$, por ser la parte de un total.
- $R^2 = 1$ implica que la varianza residual es nula y por lo tanto el ajuste es perfecto. En consecuencia, la relación entre ambas variables es lineal.
- $R^2 = 0$ implica que la varianza residual es igual a la varianza de la variable Y y que la variable explicativa no aporta información válida para la estimación de la variable explicada. En consecuencia, no existe relación lineal entre las dos variables.
- Cuanto más próximo a 1 esté R^2 mejor será la bondad o calidad del ajuste.

Por otra parte, en una regresión lineal se puede demostrar que $R^2 = \frac{S_R^2}{S_y^2} = \frac{S_{XY}^2}{S_x^2 S_y^2}$

que evidentemente coincide con el cuadrado del coeficiente de correlación lineal y justifica todas las propiedades antes mencionadas de ambos coeficientes.

$$r_{XY}^2 = R^2$$

Predicciones. Usos y abusos

El primer objetivo de la regresión lineal era poner de manifiesto la relación existente entre dos variables estadísticas. Una vez se constata que la hay, y se calcula la recta de regresión apropiada, esta se puede usar para obtener valores de la variable explicada, a partir de valores de la variable explicativa.

Por ejemplo, si se comprueba una buena correlación lineal entre las variables X = «horas de estudio semanal» y Y = «nota del examen», con una recta de regresión (de Y sobre X) igual a Y = 0.9 + 0.6 x se puede plantear la siguiente pregunta:

¿Qué nota puede obtener (según los datos) un alumno que estudia 10 horas semanales?

Y la respuesta es tan sencilla como calcular Y, sustituyendo en la ecuación de la recta x = 10, resultando Y = 6.9. El coeficiente determinación es el dato que indicará si la predicción obtenida es *fiable* o no, ya que es el coeficiente que informa sobre la calidad del ajuste.

En el momento de hacer predicciones hay que tener ciertas precauciones, porque es posible obtener resultados absurdos. Según la recta de regresión anterior, un alumno que estudie 20 horas por semana (x = 20) tendría un resultado de 12.9 puntos en su examen, lo que no tiene sentido si se evalúa sobre 10. La limitación de la predicción consiste en el hecho de que *solo se puede realizar para valores de X que estén situados dentro del rango de los valores de X* .

Objetivos

Los problemas deben permitir que los alumnos alcancen los objetivos didácticos:

2*a*) Saber analizar y extraer información de una distribución de datos bidimensional a partir de la construcción de la tabla de doble entrada.
2*b*) Saber extraer conclusiones del análisis, tanto de las distribuciones marginales como de las condicionadas de una distribución de datos bidimensional.
2*c*) Distinguir gráfica y analíticamente si las dos variables de una distribución de datos bidimensional tienen relación lineal.
2*d*) Saber calcular e interpretar la covarianza, así como aplicar las propiedades que este estadístico cumple.
2*e*) Saber calcular el coeficiente de correlación lineal así como su interpretación.
2*f*) Construir la recta de regresión lineal de una variable estadística respecto a la otra en una distribución de datos bidimensional.
2*g*) En una distribución de datos bidimensional, saber predecir el valor de una variable a partir de un valor de la otra mediante la recta de regresión y conocer su fiabilidad.

La tabla siguiente nos muestra cómo están distribuidos los objetivos según los Ejercicios:

OBJETIVOS \ EJERCICIOS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
2a	X	X	X	X															
2b	X	X	X	X															
2c					X	X	X												
2d						X	X												
2e								X	X	X	X	X	X	X	X	X	X	X	X
2f											X	X	X	X	X	X	X	X	X
2g											X	X	X	X	X	X	X	X	X

Enunciados

2*a*) Saber analizar y extraer información de una distribución de datos bidimensional a partir de la construcción de la tabla de doble entrada.
2*b*) Saber extraer conclusiones del análisis, tanto de las distribuciones marginales como de las condicionadas de una distribución de datos bidimensional

Ejercicio 1

Una empresa ha entrevistado a veinticinco de sus trabajadores con tareas administrativas para conocer el grado de implicación en su formación profesional. A cada uno le preguntó el número de cursos de formación de más de 30 horas y el número de cursos de perfeccionamiento de idiomas que había realizado en los últimos tres años. Los resultados son los que se muestran en la tabla siguiente:

Formación	8	9	4	5	6	7	7	9	10	7	5	6	7
Idiomas	8	8	3	5	7	7	8	10	10	7	6	7	8

Formación	8	5	8	9	8	8	7	7	9	9	8	7
Idiomas	7	5	8	8	7	8	7	7	8	10	8	8

a) Construye la tabla de frecuencias conjunta.
b) Calcula el número medio de cursos formación y el número medio de cursos de idiomas que han realizado los trabajadores de la empresa.
c) Calcula el número medio de cursos de formación que han hecho aquellos trabajadores que hicieron siete de perfeccionamiento de idiomas.
d) ¿Qué proporción de trabajadores ha realizado más de cinco cursos en ambas categorías? ¿Qué proporción de trabajadores ha hecho más de cinco cursos de formación? ¿Y más de cinco cursos de idiomas?
e) ¿Qué proporción de trabajadores ha realizado más de siete cursos de formación y más de 8 en idiomas?
f) ¿Qué porcentaje de los trabajadores que han hecho cinco o más cursos de formación, ha hecho siete o más cursos de idiomas?

Ejercicio 2

Una empresa quiere abrir un punto de venta en un barrio de una gran ciudad de la Comunidad Valenciana. Como el segmento de población al que va dirigido el producto es a personas de edades comprendidas entre 45 y 55 años, ha decidido encuestar a una muestra de 50 vecinos del barrio cuya edad está en esta franja. La tabla siguiente muestra dos de las preguntas que aparecían en la encuesta: edad e ingresos mensuales en miles de euros.

Edad	50	51	53	50	51	48	50	49	52	52	49	50	52	51	52	49	50
Ingresos mensuales	3.2	4.1	4.5	3	3.6	2.9	3.8	3.8	3.6	3.9	3	3.8	4.1	3.5	4.0	3.1	3.1

Edad	51	50	51	52	53	52	52	51	50	51	54	50	51	51	51	52	51
Ingresos mensuales	4.3	3.3	3.9	3.7	4.1	4.2	3.5	3.8	3.6	3.4	4.6	3.5	3.6	3.1	4	3.8	4.2

Edad	52	51	50	51	49	51	48	50	52	53	52	50	52	51	51	51
Ingresos mensuales	4	4.4	3.9	3.7	3.4	3.3	2.7	3.4	3.6	4.4	4.3	3.3	4.2	4.2	3.3	3.7

a) Construye la tabla de doble entrada agrupando los ingresos mensuales en intervalos de amplitud 0,5 y de manera que el extremo pequeño de la primera clase sea 2,5.

b) ¿Qué ingresos medios tienen los encuestados de 51 años? ¿Qué porcentaje de estos tiene unos ingresos inferiores a 4000 €?

c) ¿Cuál es la media de edad de los encuestados que tienen unos ingresos entre 3500 y 4000 euros? ¿Qué porcentaje de estos tienen 50 o 51 años?

d) ¿Qué porcentaje de los clientes ingresan mensualmente 4000 euros o más y tienen más de 50 años?

e) ¿Qué porcentaje de las personas encuestadas tienen más de 51 años o unos ingresos de 4000 € o más?

Ejercicio 3

El Departamento de Recursos Humanos de una empresa ha decidido realizar dos tests para seleccionar a las personas que deberán hacerse cargo de un proyecto de innovación. Las notas obtenidas por los aspirantes se muestran en la siguiente tabla:

TEST 1	7	6	5	4	5	8	7	8	9	6	5	8	6	8	7	8	7	6	6	9
TEST 2	8	7	6	6	7	10	9	9	10	8	6	10	8	9	8	8	7	8	6	8

a) Construye la tabla de doble entrada.
b) Calcula la nota media en el test 2 de los aspirantes que han obtenido un 6 en el test 1.
c) Calcula el porcentaje de aspirantes que obtienen un nota inferior a 8 en el test 2 entre aquellos que obtienen un nota en el test 1 superior a 6.

Ejercicio 4

La siguiente tabla muestra el número de personas ocupadas distribuidas atendiendo al sueldo neto de la actividad principal que desarrollan (en centenas de euro) y la edad en el año 2010, según datos recogidos del Ministerio de Trabajo y de Inmigración.

	SUELDO						
EDAD	[0, 6)	[6,10)	***[10,12)***	***[12,16)***	***[16,21)***	***[21,30)***	***[30, 40)***[1]
[16,25)	289,79	490,44	*249,08*	126,47	38,03	*1,70*	0
[25,30)	232,55	673,68	*571,85*	430,16	192,86	*20,80*	11,01
[30,45)	566,18	1777,07	*1671,91*	2190,02	1248,87	*736,77*	155,06
[45,55)	323,65	797,11	*881,81*	1123,69	724,93	*448,78*	138,99
[55,65[2]***)***	185,20	430,59	*503,77*	568,69	306,20	*225,13*	123,53

a) Construye las tablas de frecuencia de las distribuciones de las variables marginales y calcula la media aritmética de cada una.
b) Construye la tabla de frecuencia de la edad de aquellas personas ocupadas que tienen un sueldo de 1200 a 1600 euros. Calcula también la edad media de las personas que cobran entre 1200 y 1600 euros.
c) Construye la tabla de frecuencia del sueldo de aquellas personas ocupadas que tienen 30 años o más. ¿Qué sueldo medio cobran?
d) ¿Qué porcentaje de personas ocupadas tienen 45 años o más y cobran 1600 euros o más?
e) ¿Qué porcentaje de personas ocupadas tienen 45 años o más o cobran 1600 euros o más?
f) ¿Qué porcentaje de ocupados tiene menos de 30 años de aquellos que cobran 1200 euros o más?

1. En la tabla original el último intervalo es 3000 euros o más. Se ha cerrado el intervalo para hacer el ejercicio.
2. En la tabla original el último intervalo es 55 años o más. Se ha cerrado el intervalo para hacer el ejercicio.

2*c*) Distinguir gráfica y analíticamente si las dos variables de una distribución de datos bidimensional tienen relación lineal.

Ejercicio 5

La siguiente tabla muestra la población en edad de trabajar analfabeta en la Comunidad Valenciana, Madrid, Andalucía y el País Vasco a lo largo de los años 2000-2010 en miles de personas.

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Andalucía	332,28	280,99	342,14	307,41	294,91	247,90	262,25	278,03	294,34	290,37	279,91
Madrid	73,15	58,88	74,06	83,33	71,83	47,35	43,35	40,34	46,68	56,98	72,21
País Vasco	17,69	13,77	14,67	12,56	13,27	9,37	11,34	12,50	13,66	14,23	11,10
C. Valenciana	128,01	96,83	110,81	117,30	114,73	69,46	79,91	79,01	92,45	99,06	85,76

Fuente: INE

a) Representa la nube de puntos entre las variables: Población analfabeta en edad de trabajar en Andalucía y Población en edad de trabajar en la Comunidad Valenciana. ¿Qué observas en cuanto a la existencia o no de la relación lineal entre las dos variables?
b) Representa la nube de puntos entre las variables: Población analfabeta en edad de trabajar en la Comunidad de Madrid y en el País Vasco. ¿Qué observas en cuanto a la existencia o no de la relación lineal entre las dos variables?
c) Calcula el estadístico adecuado para confirmar las suposiciones que has hecho en los dos apartados anteriores.

2*c*) Distinguir gráfica y analíticamente si las dos variables de una distribución de datos bidimensional tienen relación lineal.
2*d*) Saber calcular e interpretar la covarianza, así como aplicar las propiedades que este estadístico cumple.

Ejercicio 6

Se recogieron los valores mensuales de los gastos en publicidad de una compañía ferroviaria y el número de pasajeros a lo largo de 15 meses. Los datos los muestra la tabla:

Publicidad (en miles)	10	12	8	17	10	15	10	14	19	10	11	13	16	10	12
Pasajeros (en miles)	15	17	13	23	16	21	14	20	24	17	16	18	23	15	16

a) Calcula el gasto medio y el número medio de pasajeros.
b) Haz la nube de puntos y calcula la covarianza. ¿Es coherente el valor del estadístico con la nube de puntos?
c) Si para los 15 meses posteriores se prevé que la inversión en publicidad de cada mes aumente un 10 % respecto al mismo mes del período anterior, y también se prevé que este hecho provocará un aumento del 8 % en el número de pasajeros cada mes, ¿cuál será la covarianza en este segundo período?

Ejercicio 7

Una empresa ha realizado dos tests psicotécnicos a los 9 trabajadores de un departamento como parte del proceso de selección del nuevo director del departamento. La siguiente tabla muestra los resultados obtenidos por los aspirantes:

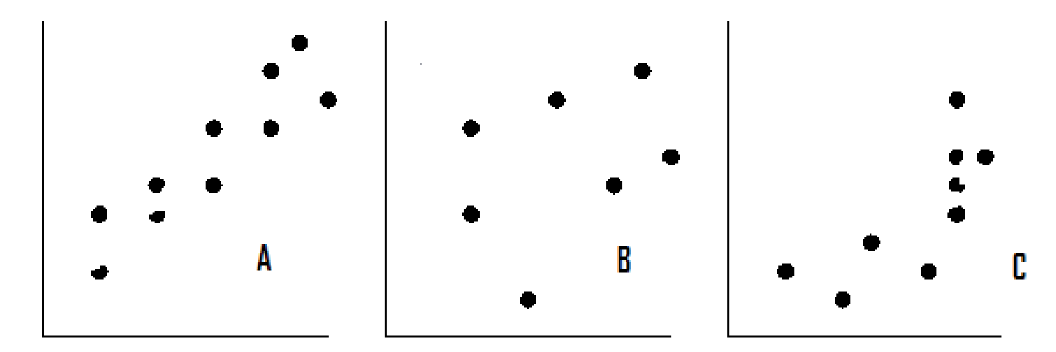
TEST 1	5	7	6	9	3	1	2	4	6
TEST 2	6	5	8	6	4	2	1	3	7

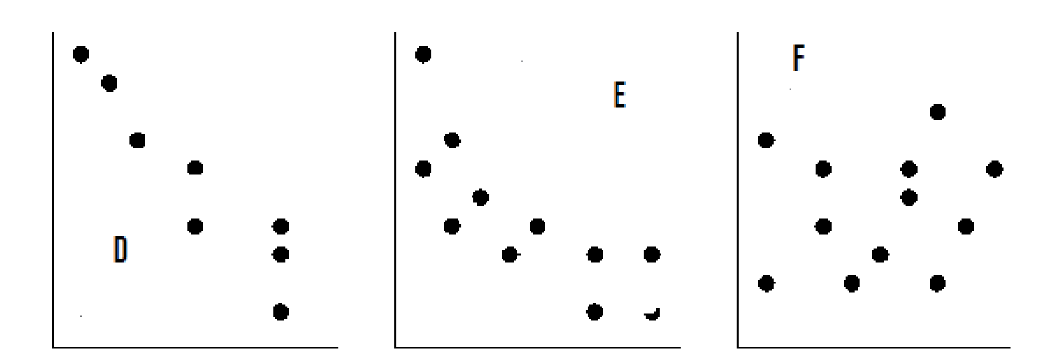
a) Calcula la covarianza. ¿Existe algún tipo de relación lineal entre las dos variables?
b) Ha habido un error en una pregunta de cada test y el tribunal decide aumentar un 5 % la puntuación de cada participante. Calcula nuevamente la covarianza.

2*e*) Saber calcular el coeficiente de correlación lineal así como su interpretación.

Ejercicio 8

Dados las siguientes nubes de puntos, contesta:





a) Asocia cada nube de puntos con el valor del coeficiente de correlación que le corresponde entre estos valores: –0,9; 0,4; 0,95: –0,65; 0,1; 0,6. Razona la respuesta.
b) Indica para cada nube de puntos el signo de la covarianza y di cuál es su significado.

Ejercicio 9

La siguiente tabla muestra el gasto total promedio, el gasto medio en alimentos y bebidas no alcohólicas y el gasto en vivienda, agua, electricidad, gas y otros combustibles en euros, por número de personas que forman la unidad familiar en el año 2009,[3] según datos del INE.

Número de miembros de la familia	*1*	*2*	*3*	*4*	*5*	*6 o más*
Gastos medios totales	18355,25	27755,08	33414,09	38576,14	40699,09	41562,31
Gastos en vivienda, agua, electricidad, gas y otros combustibles	7493,88	8990,72	9205,13	9645,19	10114,49	9272,18

a) ¿Existe una fuerte relación lineal entre el número de miembros que viven en un hogar y el gasto medio total? Razona la respuesta.
b) ¿Y entre el número de miembros que viven en un hogar y el gasto en vivienda, agua, electricidad, gas y otros combustibles? Razona la respuesta.

3. Para realizar el ejercicio, considera 7 miembros en el intervalo *6 o más*.

Ejercicio 10

El director de Recursos Humanos de una empresa ha realizado dos tests psicotécnicos para seleccionar a las personas que deben trabajar en el Departamento de Marketing. Se han presentado 9 personas y los resultados obtenidos en cada uno de los tests han sido los siguientes:

TEST 1	175	181	192	211	235	255	275	286	292
TEST 2	169	185	202	219	240	266	295	329	357

Teniendo en cuenta los resultados de los tests, ¿crees que el director podría haber eliminado uno de los dos tests para decidir los candidatos? Razona la respuesta.

2*f*) Saber calcular el coeficiente de correlación lineal así como su interpretación.
2*e*) Construir la recta de regresión lineal de una variable estadística respecto a la otra en una distribución de datos bidimensional.
2*g*) En una distribución de datos bidimensional, saber predecir el valor de una variable a partir de un valor de la otra mediante la recta de regresión y conocer su fiabilidad.

Ejercicio 11

En una muestra de 150 empresas del sector de servicios se recogen datos sobre el número de trabajadores de la empresa (X) y la facturación (Y) anual en millones de euros. Los resultados se muestran resumidos en los siguientes estadísticos:

$$\overline{X} = 14 \text{ trabajadores} \quad \overline{Y} = 100 \text{ millones} \quad S_X = 2 \text{ trabajadores} \quad S_Y = 25 \text{ n}$$

$$S_{XY} = 45 \text{ trabajadores} \times \text{millón}$$

a) Calcula el coeficiente de correlación lineal e interprétalo.
b) Calcula el modelo de regresión lineal que mejor aproxima la facturación en función del número de trabajadores
c) En función de este ajuste, calcula de forma aproximada la cantidad que se espera que facture una empresa con 15 trabajadores. ¿Es fiable esta predicción? Razona la respuesta.
d) Calcula el modelo de regresión lineal que mejor aproxima el número de trabajadores en función de la facturación.
f) En función de este ajuste calcula de forma aproximada el número de trabajadores que se espera que tenga una empresa que facture 105 millones. ¿Es fiable esta predicción? Razona la respuesta.

Ejercicio 12

Las dos tablas siguientes muestran el grado medio de satisfacción de los ocupados según el trabajo que realizan por edad y por el nivel de estudios en 2010. Los datos han sido extraídos del Ministerio de Trabajo e Inmigración.

NIVEL ESTUDIOS	*GRADO DE SATISFACCIÓN*
1	7,05
2	7,09
3	7,21
4	7,23
5	7,50
6	7,55

EDAD	*GRADO DE SATISFACCIÓN*
[16,25)	7,33
[25,30)	7,39
[30,45)	7,37
[45,55)	7,30
[55,65)[4]	7,43

Hay que decir que la variable nivel de estudios ha sido convertida a numérica discreta para ser graduable. Así, la equivalencia es: 1 = menos que Primarios; 2 = Primarios; 3 = Secundarios; 4 = Bachillerato; 5 = Formación Profesional y 6 = Universitarios. Esta conversión se ha hecho a efectos didácticos:

a) Calcula el coeficiente de relación lineal de ambos pares de variables. ¿En cuál de las dos convendría calcular la recta de regresión?

b) Calcula la recta de regresión del grado de satisfacción en función del nivel de estudios.

Ejercicio 13

El grado medio de satisfacción medio de los ocupados, según el trabajo que realizan por nivel de ingresos y por sexo en el año 2010, se muestra en la tabla siguiente. Los datos han sido extraídos del Ministerio de Trabajo e Inmigración.

NIVEL DE INGRESOS	*GRADO DE SATISFACIÓN HOMBRES*	*GRADO DE SATISFACIÓN MUJERES*
[0,600)	6,19	7,253
[600,1000)	6,83	7,234
[1000,1200)	7,28	7,339
[1200,1600)	7,39	7,61
[1600,2100)	7,60	7,768
[2100,3000)	7,82	7,682
[3000,4000)[5]	7,925	7,499

4. En la tabla original el último intervalo es 55 años o más. Se ha cerrado el intervalo para poder hacer el ejercicio.
5. En la tabla original el último intervalo es 3000 euros o más. Se ha cerrado el intervalo para poder hacer el ejercicio.

a) Calcula el coeficiente de correlación lineal entre las variables Nivel de ingresos y Grado de satisfacción en los hombres y entre las variables Nivel de ingresos y Grado de satisfacción en las mujeres. ¿Qué conclusiones se pueden obtener?

b) Calcula la recta de regresión que explique el grado de satisfacción medio en el trabajo de los hombres en función del nivel de ingresos.

Ejercicio 14

El número total de expedientes de regulación del trabajo a lo largo de los años 2001-2010, según los datos extraídos del Ministerio de Trabajo e Inmigración, son los que se muestran en la tabla.

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Alicante	139	169	224	268	292	180	164	393	939	679
Castellón	49	59	55	76	88	59	58	291	939	777

a) ¿Existe algún tipo de relación lineal entre las variables? ¿Es fuerte esta relación? Razona las respuestas.

b) Calcula la recta de regresión lineal que relaciona el número de expedientes totales en Castellón en función de los de Alicante.

Ejercicio 15

La siguiente tabla muestra el número total de hipotecas firmadas, así como la tasa de paro en España en el período 2004-2010, según datos extraídos del INE.

	Hipotecas	***Tasa de paro***
2004	1608497	8,1
2005	1798630	9,2
2006	1896515	8,3
2007	1780627	8,6
2008	1283374	13,9
2009	1082587	18,83
2010	961601	20,05

a) ¿Existe algún tipode relación lineal entre las variables? ¿Es fuerte esta relación? Razona las respuestas.

b) Calcula la recta de regresión lineal que relaciona el número hipotecas firmadas en función de la tasa de desempleo.

Ejercicio 16

La siguiente tabla muestra el número de horas extraordinarias totales en miles (remuneradas y no remuneradas) realizadas en el conjunto de España, así como las tasas de paro desde el primer trimestre de 2008 hasta el último trimeste del año 2010. Los datos han sido extraídos del INE.

Trimestres	*Número total de horas extras*	*Tasa de paro*
2010TIV	5574,9	20,33
2010TIII	5058,9	19,79
2010TII	6002,7	20,09
2010TI	6154,1	20,05
2009TIV	6493,2	18,83
2009TIII	6069	17,93
2009TII	7042	17,92
2008TIV	8398,4	13,91
2008TIII	8813,2	11,33
2008TII	9794,4	10,44
2008TI	10058,1	9,63

a) Halla, en su caso, las recta de regresión que explica el número de horas extras en función de la tasa de desempleo.

b) En el primer trimestre de 2009 la tasa de paro era del 17,36 %. Da una estimación del número de horas extras en este trimestre, así como una medida de su fiabilidad.

Ejercicio 17

El número total de expedientes de regulación del trabajo a lo largo de los años 2001-2010 en Cataluña y la Comunidad Valenciana, extraídos del Ministerio de Trabajo e Inmigración, son las que se muestran en la tabla, a excepción de los datos del año 2005 que se han omitido.

Año	*Cataluña*	*Comunidad Valenciana*
2001	661	465
2002	724	494
2003	608	594
2004	565	619
2006	455	413
2007	470	487
2008	874	1286
2009	3964	3490
2010	3318	2810

Se sabe que el número de expedientes en 2005 en Cataluña fue 512. Da una estimación, si es conveniente, del número de expedientes en la Comunidad Valenciana así como una medida del ajuste.

Ejercicio 18

En un museo se desea estudiar la repercusión que tienen las quejas realizadas por los visitantes y los ingresos. Para realizarlo, se observaron las dos variables a lo largo de las últimas diez semanas. Las visitas están expresadas en decenas de asistentes.

Quejas	18	26	30	33	38	39	42	44	46	49
Visitas	107	105,5	105	104,4	104,3	104	103,7	103,4	103,1	103

Si la entrada al museo tiene un coste de 3,6 euros, estima los ingresos del museo si en una semana se hubieran producido 43 quejas.

Ejercicio 19

La siguiente tabla muestra el número de personas ocupadas distribuidas atendiendo el sueldo neto de la actividad principal que desarrollan (en centenas de euro) y el nivel de estudios que tenían en 2010, según datos recogidos del Ministerio de Trabajo y de Inmigración. Hay que decir, sin embargo, que la variable nivel de estudios ha sido convertida a numérica discreta para ser graduable. Así, la equivalencia es: 1 = menos que Primarios; 2 = Primarios; 3 = Secundarios; 4 = Bachillerato; 5 = Formación Profesional y 6 = Universitarios. Esta conversión se ha hecho a efectos didácticos.

Nivel de estudios	SUELDO [0, 6)	[6,10)	[10,12)	[12,16)	[16,21)	[21,30)	[30, 40)
1	8,75	21,67	*15,00*	7,93	5,65	*0,74*	1,56
2	293,18	790,92	*601,61*	472,64	92,82	*56,30*	3,77
3	538,08	1551,52	*1226,20*	1098,34	340,72	*74,78*	13,56
4	323,39	670,29	*607,31*	709,62	313,35	*142,00*	53,20
5	303,28	801,87	843,80	982,42	444,90	183,90	50,73
6	164,47	439,26	619,51	1155,75	1230,07	919,40	282,14

a) ¿Están relacionadas linealmente el sueldo y el nivel de estudios?
b) Calcula una estimación del sueldo que cobraría una persona ocupada que tuviera un nivel de estudios equivalente a 4,5, así como su fiabilidad.

Ayudas

En este apartado se presentarán las ayudas para emplear en caso de ser necesario a la hora de realizar los ejercicios y problemas. Es conveniente no hacer un abuso excesivo de estas ayudas, es decir, antes de emplearlas hay que pensar el problema al menos durante unos 10-15 minutos. Después se consultará la ayuda de tipo 1 y se intentará resolver el ejercicio con esta ayuda. Si no es posible resolverlo, entonces se consultará la ayuda de tipo 2, y en último término la solución.

Ayudas Tipo 1

Ejercicio 1

Consulta la introducción teórica.

Ejercicio 2

Consultar la introducción teórica. En el apartado *e*) nota que pide el porcentaje de las personas encuestadas que tienen más de 51 años o unos ingresos de 4000 euros o más. Hay que contar pues el número de personas que cumplen una condición o la otra.

Ejercicio 3

Consulta la introducción teórica.

Ejercicio 4

Consulta la introducción teórica.

En el apartado *c*) debes tener presente que la variable que condiciona, en este caso la edad, incluye más de un intervalo. Entonces hay que agrupar las frecuencias conjuntas adecuadamente.

Ejercicio 5

Consulta la introducción teórica.

En el apartado *c*) tienes que calcular el estadístico que permite contrastar si dos variables estadísticas están relacionadas linealmente.

Ejercicio 6

Para hacer el gráfico tienes que seguir las indicaciones del ejercicio 5. Para afirmar que el gráfico es coherente con el resultado de la covarianza debes fijarte en el signo del estadístico.

Para responder a la aparta *c*) hay que aplicar una propiedad.

Ejercicio 7

Consulta la introducción teórica.

Ejercicio 8

El coeficiente de correlación lineal informa de las mismas cosas que lo hace la covarianza. Puedes consultar la introducción teórica para saber las propiedades.



Ejercicio 9

Hay que calcular un estadístico que mida el grado de relación lineal entre dos variables.

Ejercicio 10

El director podría haber eliminado una de las dos pruebas siempre y cuando las dos discriminen a las mismas personas.

Ejercicio 11

El apartado *a*) es directo y para el resto de apartados hay que construir las funciones lineales que mejoran el ajuste. Para hacer las predicciones hay que sustituir las valores de las variables explicativas en las fórmulas.

Ejercicio 12

Hay que hacer lo mismo que en el ejercicio 11 pero en este caso tienes que calcular los estadísticos a partir de los datos. Únicamente habrá que calcular la recta de regresión que tenga un coeficiente de correlación superior a 0,8.

Ejercicio 13

Véase la ayuda del ejercicio 12.

Ejercicio 14

Para contestar la pregunta *a*) tienes que calcular el estadístico que informa sobre el grado de relación lineal entre dos variables.

Ejercicio 15

Para contestar la pregunta *a*) tienes que calcular el estadístico que informa sobre el grado de relación lineal entre dos variables.

Ejercicio 16

Únicamente cuando sea pertinente convendrá calcular la recta y hacer la predicción.

Ejercicio 17

Únicamente cuando sea pertinente convendrá calcular la recta y hacer la predicción.

Ejercicio 18

Debes tener en cuenta que los ingresos dependen completamente del número de visitas. Es decir, si sabes un estimación de las visitas sabrás una estimación de los ingresos.

Ejercicio 19

Para contestar la pregunta del apartado *a*) tienes que calcular el coeficiente de correlación. Utiliza la tabla para hacer los cálculos de los estadísticos y los productos que se necesitan para calcular la covarianza.

Ayudas Tipo 2

Ejercicio 1

Para contestar la pregunta *b*) hay que construir la tabla de la distribución marginal, y para la pregunta *c*), la tabla de la distribución condicionada:

Cursos de formación	n_i	$x_i \cdot n_i$
4	1	4
5	3	15
6	2	12
7	7	49
8	6	48
9	5	45
10	1	10
	25	183

X/(Y = 7)	n_i	$x_i \cdot n_i$
6	2	12
7	4	28
8	2	16
	8	**56**

Para hacer los recuentos puedes ayudarte de la tabla de doble entrada:

Y = cursos de idiomas

X = cursos de formación

X \ Y	3	5	6	7	8	10	ni.
4	1						1
5		2	1				3
6				2			2
7				4	3		7
8				2	4		6
9					3	**2**	5
10						**1**	1
n.j	1	2	1	8	10	3	25

Ejercicio 2

En el apartado *e*) nota que pide el porcentaje de las personas encuestadas que tienen más de 51 años o unos ingresos de 4000 € o más. Observa que este número pedido es igual a: (encuestados que ingresan más de 4000) + (encuestados que tienen más de 51 años) – (encuestados que tienen más de 51 años e ingresan más de 4000 €).

Ejercicio 3

Como es muy similar al ejercicio 1, consúltalo en sus ayudas.

Ejercicio 4

Muy similar al ejercicio 1. Consulta en las ayudas de este ejercicio. Calcula previamente las marcas de las clases de cada variable. En el apartado *c*) puedes utilizar la siguiente tabla:

EDAD	[0, 6)	[6,10)	[10,12)	[12,16)	[16,21)	[21,30)	[30,40)
[30,45)	566,18	1777,07	*1671,91*	2190,02	1248,87	*736,77*	155,06
[45,55)	323,65	797,11	*881,81*	1123,69	724,93	*448,78*	138,99
[55,65)	185,2	430,59	*503,77*	568,69	306,2	*225,13*	123,53
n.j	**1075,03**	**3004,77**	**3057,49**	**3882,4**	**2280**	**1410,68**	**417,58**



Ejercicio 5

En el apartado *c*) has de calcular la covarianza.

Ejercicio 6

Para responder en el apartado *c*) debes aplicar la propiedad de la covarianza, teniendo en cuenta que hay que definir dos variables nuevas a partir de las dos anteriores: X' = gastos en el segundo período y Y' = número de pasajeros en el segundo período. Según el enunciado X' = 1,1 · X y Y' = 1,08 · Y.

Ejercicio 7

Misma ayuda que en el ejercicio 6. En este caso las nuevas variables son: X' = nota del test 1 tras el aumento y Y' = nota del test 2 después del aumento.
Según el enunciado X' = 1,05 · X y Y' = 1,058 · Y.

Ejercicio 8

Se remite a la ayuda de tipo 1 por ser lo suficientemente aclaratoria.

Ejercicio 9

Hay que calcular el coeficiente de correlación lineal.

Ejercicio 10

Hay una relación casi funcional entre las dos variables. El director podría haber eliminado una de las dos pruebas siempre y cuando exista.

Ejercicio 11

Las funciones lineales que hay que calcular son las rectas de regresión. Para saber la fiabilidad de las predicciones busca el coeficiente de determinación.

(cc) P. Juan Verdoy / M. J. Beltrán / M. J. Peris - ISBN: 978-84-15444-38-1 Problemas resueltos de estadística aplicada a las ciencias sociales - UJI - DOI: http://dx.doi.org/10.6035/Sapientia100



Ejercicio 12

Se remite a la ayuda de tipo 1 por ser lo suficientemente aclaratoria.

Ejercicio 13

Véase la ayuda del ejercicio 12.

Ejercicio 14

Para contestar la pregunta del apartado *a*) tienes que calcular el coeficiente de correlación.

Ejercicio 15

Para contestar la pregunta del apartado *a*) tienes que calcular el coeficiente de correlación.

Ejercicio 16

Únicamente si el coeficiente de correlación es cercano a 1 o a –1 es pertinente calcular la recta y hacer la predicción.

Ejercicio 17

Únicamente si el coeficiente de correlación es cercano a 1 o a –1 es pertinente calcular la recta y hacer la predicción.

Ejercicio 18

Si existe tipo de relación lineal entre el número de quejas y el número de visitas, entonces podremos encontrar la recta de regresión entre el número de visitas y el de quejas y, con posterioridad, se podrán estimar los ingresos.

Ejercicio 19

Se remite a la ayuda de tipo 1 por ser lo suficientemente aclaratoria.

P. Juan Verdoy / M. J. Beltrán / M. J. Peris - ISBN: 978-84-15444-38-1 Problemas resueltos de estadística aplicada a las ciencias sociales - UJI - DOI: http://dx.doi.org/10.6035/Sapientia100

Índice

Soluciones

Ejercicio 1

Una empresa ha entrevistado a veinte y cinco de sus trabajadores con tareas administrativas para conocer el grado de implicación en su formación profesional. A cada uno se le preguntó el número de cursos de formación de más de 30 horas y el número de cursos de perfeccionamiento de idiomas que había realizado en los últimos tres años. Los resultados son los que se muestran en la tabla siguiente:

Formación *EST*	8	9	4	5	6	7	7	9	10	7	5	6	7
Idiomas *ECO*	8	8	3	5	7	7	8	10	10	7	6	7	8

Formación *EST*	8	5	8	9	8	8	7	7	9	9	8	7
Idiomas *ECO*	7	5	8	8	7	8	7	7	8	10	8	8

a) Construye la tabla de frecuencias conjunta.
b) Calcula el número medio de cursos de formación y el número medio de cursos de idiomas que han realizado los trabajadores de la empresa.
c) Calcula el número medio de cursos de formación que han hecho aquellos trabajadores que hicieron siete de perfeccionamiento de los idiomas.
d) ¿Qué proporción de trabajadores ha realizado más de cinco cursos en ambas categorías? ¿Qué proporción de trabajadores ha hecho más de cinco cursos de formación? ¿Y más de cinco cursos de idiomas?
e) ¿Qué proporción de trabajadores ha realizado más de siete cursos de formación y más de ocho en idiomas?
f) ¿Qué porcentaje de los trabajadores que han hecho cinco cursos o más de formación, ha hecho siete cursos o más de idiomas?

Solución

a) Construye la tabla de frecuencias conjunta.

Para construir la tabla de doble entrada, en primer lugar, hay que ordenar los datos de las dos variables de menor a mayor y construir una cuadrícula, por lo que en la primera fila se sitúan las diferentes categorías o valores que toma una de las variables, y en la primera columna los valores o las categorías relativas a la segunda. Así, el número que hay que asociar a cada celda de la tabla de doble entrada es la frecuencia absoluta conjunta del dato bivariante, formada por los valores correspondientes ubicados en la primera fila y en la primera columna.

En nuestro caso podemos representar en las filas la variable Cursos de formación y en las columnas la variable Cursos de idiomas. Así pues:

X = Cursos de formación \ Y = Cursos de idiomas	3	5	6	7	8	10	ni.
4	1						1
5		2	1				3
6				2			2
7				4	3		7
8				2	4		6
9					3	2	5
10						1	**1**
n.j	**1**	**2**	**1**	**8**	**10**	**3**	**25**

b) Calcula el número medio de cursos de formación y el número medio de cursos de idiomas que han realizado los trabajadores de la empresa.

Para contestar estas preguntas hay que estudiar las dos distribuciones marginales; X = Cursos de formación y Y = Cursos de idiomas. Es decir: variable Cursos de formación en las filas y en las columnas la variable Cursos de idiomas. Así pues:

Cursos de formación	n_i	$x_i \cdot n_i$
4	1	4
5	3	15
6	2	12
7	7	49
8	6	48
9	5	45
10	1	10
	25	183

Cursos de idiomas	n_i	$y_i \cdot n_i$
3	1	3
5	2	10
6	1	6
7	8	56
8	10	80
10	3	30
	25	185

$$\bar{x} = \frac{183}{25} = 7{,}32 \text{ cursos} \qquad \bar{y} = \frac{185}{25} = 7{,}4 \text{ cursos}$$

c) Calcula el número medio de cursos de formación que han hecho aquellos trabajadores que han realizado siete de perfeccionamiento de los idiomas.

En primer lugar, en este apartado, hay que construir la variable cursos de formación condicionada a que el valor de la variable cursos de idiomas es 7. Es decir, hay que construir la variable X / (Y = 7). Extrayendo de la tabla este distribución marginal:

X/(Y = 7)	n_i	$x_i \cdot n_i$
6	2	12
7	4	28
8	2	16
	8	**56**

En consecuencia:

$$\overline{X/(Y=7)} = \frac{56}{8} = 7 \text{ cursos}$$

d) ¿Qué proporción de trabajadores ha realizado más de cinco cursos en ambas categorías? ¿Qué proporción de trabajadores ha hecho más de cinco cursos de formación? ¿Y más de cinco cursos de idiomas?

Y = Cursos de idiomas

X = Cursos de formación	3	5	6	7	8	10	ni.
4	1						1
5		2	1				3
6				2			2
7				4	3		7
8				2	4		6
9					3	**2**	5
10						**1**	1
n.j	1	2	1	8	10	3	25

Para contestar a la primera pregunta, es necesario que contemos el número de datos que cumplen las dos condiciones a la vez, es decir, los datos que aparecen con la celda en color rojo:

Así pues, como hay 17, el porcentaje solicitado es $\frac{17}{25} = \frac{0,68}{1} = \frac{68}{100}$, un 68 %.

De esta forma, se cuenta el número de personas que ha hecho más de 5 cursos de formación, que son 17, lo que representa un 68 %. También podemos calcular fácilmente el número de trabajadores que ha hecho más de 5 cursos de idiomas, que son 18; los cuales representan un 72 %.

e) ¿Qué proporción de trabajadores han realizado más de 7 cursos de formación y más de 8 en idiomas?

En este apartado hay que contar el número de trabajadores que cumplen ambas condiciones. En este caso son 3 (las que aparecen en amarillo en el apartado *d*)). Por tanto, el porcentaje es de 3 de 25, un 12 %.

f) ¿Qué porcentaje de los trabajadores que ha hecho cinco cursos o más de formación, ha hecho siete o más cursos de idiomas?

En este apartado hay que observar que cambia la cantidad total sobre la que tenemos que hacer el porcentaje. Es decir, no se nos pide el porcentaje sobre los 25 datos, sino sobre los que han hecho cinco o más cursos de formación que son 24. De estos hay 21 que han hecho 7 cursos o más de idiomas. En consecuencia, hay un 87,5 %.

Ejercicio 2

Una empresa quiere abrir un punto de venta en un barrio de una gran ciudad de la Comunidad Valenciana. Como el segmento de población al que va dirigido el producto es a personas de edades comprendidas entre 45 y 55 años, ha decidido encuestar a una muestra de 50 vecinos del barrio cuya edad está en esta franja. La tabla siguiente muestra dos de las preguntas que aparecían en la encuesta: edad e ingresos mensuales en miles de euros.

Edad	50	51	53	50	51	48	50	49	52	52	49	50	52	51	52	49	50
Ingresos mensuales	3.2	4.1	4.5	3	3.6	2.9	3.8	3.8	3.6	3.9	3	3.8	4.1	3.5	4.0	3.1	3.1

Edad	51	50	51	52	53	52	52	51	50	51	54	50	51	51	51	52	51
Ingresos mensuales	4.3	3.3	3.9	3.7	4.1	4.2	3.5	3.8	3.6	3.4	4.6	3.5	3.6	3.1	4	3.8	4.2

Edad	52	51	50	51	49	51	48	50	52	53	52	50	52	51	51	51
Ingresos mensuales	4	4.4	3.9	3.7	3.4	3.3	2.7	3.4	3.6	4.4	4.3	3.3	4.2	4.2	3.3	3.7

a) Construye la tabla de doble entrada agrupando los ingresos mensuales en intervalos de amplitud 0,5 y de manera que el extremo inferior de la primera clase sea 2,5.
b) ¿Qué ingresos medios tienen los encuestados de 51 años? ¿Qué porcentaje de estos tiene unos ingresos inferiores 4000 euros?
c) ¿Cuál es la media de edad de los encuestados que tienen unos ingresos entre 3500 y 4000 euros? ¿Qué porcentaje de estos tienen 50 o 51 años?

d) ¿Qué porcentaje de los clientes ingresan mensualmente 4000 euros o más y tienen más de 50 años?

e) ¿Qué porcentaje de las personas encuestadas tienen más de 51 años o unos ingresos de 4000 euros o más?

Solución

a) Construye la tabla de doble entrada agrupando los ingresos mensuales en intervalos de amplitud 0,5 y de manera que el extremo inferior de la primera clase sea 2,5.

En primer lugar, hay que construir los intervalos de la variable que debe estar agrupada; en este caso los ingresos mensuales. Así pues, siguiendo las indicaciones que da el enunciado, los intervalos son [2,5, 3); [2,5, 3); [3, 3,5); [3,5, 4,); [4, 4,5); [4,5, 5).

Siguiendo ahora las indicaciones del ejercicio anterior apartado *a*), podemos construir la tabla de doble entrada con las variables X = Ingresos mensuales y Y = Edad.

Y = Edad

X = Ingresos mensuales	48	49	50	51	52	53	54	ni.
[2,5 , 3)	2							2
[3, 3,5)		3	6	4				13
[3,5 , 4)		1	5	7	6			19
[4 , 4,5)				6	6	2		14
[4,5 , 5)						1	1	2
n.j	2	4	11	17	12	3	1	50

b) ¿Qué ingresos medios tienen los encuestados de 51 años? ¿Qué porcentaje de estos tiene unos ingresos inferiores 4000 euros?

Hay que construir la variable Ingresos condicionada a que la edad sea de 51 años. Es decir X / Y = 51. Además, como que la variable X está agrupada en intervalos es necesario calcular las marcas de clase para obtener la media pedida. Así, la tabla de frecuencias de esta variable es:

X/Y = 51	c_i	n_i	$c_i \cdot n_i$
[3, 3,5)	3,25	4	13
[3,5 , 4)	3,75	7	26,25
[4 , 4,5)	4,25	6	25,5
n.j		17	64,75

En consecuencia, la media será:

$$\overline{X/(Y=51)} = \frac{64,75}{17} = 3,809 \text{ miles €}$$

Por otra parte, hay 11 encuestados con un sueldo inferior 4000 euros, lo que representa un 64,7 % de 17.

c) ¿Cuál es la media de edad de los encuestados que tienen unos ingresos entre 3500 y 4000 euros? ¿Qué porcentaje de estos tienen 50 o 51 años?

Ahora hay que construir la variable Edad condicionada a unos ingresos de entre 3500 y 4000 euros. Es decir, Y / (X = [3,5, 4)). La tabla de frecuencias de la variable es:

Y/X=[3,5 , 4)	n_i	$x_i \cdot n_i$
49	1	49
50	5	250
51	7	357
52	6	312
n.j	19	968

En consecuencia, la media será:

$$\overline{Y/(X=[3{,}5\,,4)}} = \frac{968}{19} = 50{,}95 \text{ años}$$

Por otra parte, hay 12 que tienen 50 o 51 años, lo que representa un 63,16 %.

d) ¿Qué porcentaje de los clientes ingresan mensualmente 4000 euros o más y tienen más de 50 años?

Hay que hacer el recuento de las personas encuestadas que cumplen las dos condiciones que cita el enunciado. Las que se traducen en que deben ingresar más de 4000 euros (X > 4) y tener más de 50 años (Y > 50).

Revisando la tabla se observa que hay 16, los cuales representan el 32 % de los 50 encuestados.

e) ¿Qué porcentaje de las personas encuestadas tienen más de 51 años o unos ingresos de 4000 euros o más?

La pregunta es diferente a la anterior, ya que nos pregunta qué porcentaje cumple una condición o la otra. Es decir, hay que contar los encuestados que ingresan más de 4000 euros (cielo con mayor tamaño) o tienen más de 51 años (cielo rojas). Hay que notar que no podemos sumar el número de personas que cumplen una condición más el número de personas que cumplen la otra, ya que de esta manera estaríamos contando las personas que cumplen ambas condiciones dos veces.

Y = Edad

X = Ingresos mensuales	48	49	50	51	52	53	54	ni.
[2,5 , 3)	2							**2**
[3, 3,5)		3	6	4				**13**
[3,5 , 4)		1	5	7	6			**19**
[4 , 4,5)				**6**	**6**	**2**		**14**
[4,5 , 5)						**1**	**1**	2
n.j	2	4	11	17	12	3	1	50

Por tanto, el número de encuestados pedidos son: (encuestados que ingresan más de 4000) + (encuestados que tienen más de 51 años) – (encuestados que tienen más de 51 años e ingresan más de 4000 euros) = 16 + 16 – 10 = 22. Por lo tanto, el porcentaje que representan respecto a 50 es el 44 %.

Ejercicio 3

El Departamento de Recursos Humanos de una empresa ha decidido realizar dos tests para seleccionar a las personas que han de hacerse cargo de un proyecto de innovación. Las notas obtenidas por los aspirantes se muestran en la siguiente tabla:

TEST 1	7	6	5	4	5	8	7	8	9	6	5	8	6	8	7	8	7	6	6	9
TEST 2	8	7	6	6	7	10	9	9	10	8	6	10	8	9	8	8	7	8	6	8

a) Construye la tabla de doble entrada.
b) Calcula la nota media en el test 2 de los aspirantes que han obtenido un 6 en el test 1.
c) Calcula el porcentaje de aspirantes que obtienen un nota inferior a 8 en el test 2 de entre aquellos que obtienen un nota en el test 1 superior a 6.

Solución

De la misma manera que hacíamos en el apartado *a*) del primer ejercicio, construimos la tabla de doble entrada.

	Y = Test 1						
X = Test 2	**4**	**5**	**6**	**7**	**8**	**9**	**ni.**
6	1	2	1				**4**
7		1	1	1			**3**
8			3	2	1	1	**7**
9				1	2	0	**3**
10					2	1	**3**
n.j	**1**	**3**	**5**	**4**	**5**	**2**	**20**

b) Calcula la nota media en el test 2 de los aspirantes que han obtenido un 6 en el test 1.

Se nos pide que calculemos la media aritmética de la variable test 2 condicionada a que la variable Test 1 sea 6. Es decir; X / Y = 6. Calculamos la tabla de frecuencias correspondiente.

Y/X = 6	ni	xi · ni
6	1	6
7	1	7
8	3	24
n.j	5	37

En consecuencia, la media será:

$$\overline{X/(Y=6)} = \frac{37}{5} = 7,4 \ \text{ puntos}$$

c) Calcula el porcentaje de aspirantes que obtienen un nota inferior a 8 en el test 2 de entre aquellos que obtienen un nota en el test 1 superior a 6.

De la misma manera que en el apartado *f*) del ejercicio 1, en este apartado no se nos pide el porcentaje sobre las 20 aspirantes, sino sobre los que han sacado más de 6 en el Test 1, que son 11. De estas, hay 1 que ha sacado menos de 8. En consecuencia, hay un 9,09 %.

Ejercicio 4

La siguiente tabla muestra el número de personas ocupadas distribuidas atendiendo al sueldo neto de la actividad principal que desarrollan (en centenas de euro) y la edad en el año 2010, según datos recogidos del Ministerio de Trabajo y de Inmigración.

	SUELDO						
EDAD	[0, 6)	[6,10)	***[10,12)***	***[12,16)***	***[16,21)***	***[21,30)***	***[30, 40)***[6]
[16,25)	289,79	490,44	*249,08*	126,47	38,03	*1,70*	0
[25,30)	232,55	673,68	*571,85*	430,16	192,86	*20,80*	11,01
[30,45)	566,18	1777,07	*1671,91*	2190,02	1248,87	*736,77*	155,06
[45,55)	323,65	797,11	*881,81*	1123,69	724,93	*448,78*	138,99
[55,65[7])	185,20	430,59	*503,77*	568,69	306,20	*225,13*	123,53

a) Construye las tablas de frecuencia de las distribuciones de las variables marginales y calcula la media aritmética de cada una.
b) Construye la tabla de frecuencia de la edad de aquellas personas ocupadas que tienen un sueldo de 1200 a 1600 euros. Calcula también la edad media de las personas que cobran entre 1200 y 1600 euros.
c) Construye la tabla de frecuencia del sueldo de aquellas personas ocupadas que tienen 30 años o más. ¿Qué sueldo medio cobran?
d) ¿Qué porcentaje de personas ocupadas tienen años 45 o más y cobra 1600 euros o más?

6. En la tabla original el último intervalo es 3000 euros o más. Se ha cerrado el intervalo para hacer el ejercicio.
7. En la tabla original el último intervalo es 55 años o más. Se ha cerrado el intervalo para hacer el ejercicio.

e) ¿Qué porcentaje de personas ocupadas tiene años 45 o más o cobra 1600 euros o más?

f) ¿Qué porcentaje de ocupados tiene menos de 30 años de aquellos que cobran 1200 euros o más?

Solución

a) Construye las tablas de frecuencia de las distribuciones de las variables marginales y calcula la media aritmética de cada una.

De la misma manera que en el apartado *a*) del ejercicio 1, se construye las tablas de frecuencia de las variables marginales X = Edad de las personas ocupadas y Y = Sueldo de las personas ocupadas.

Edad	c_i	n_i	$c_i \cdot n_i$
[16,25)	20,5	1195,51	24507,955
[25,30)	27,5	2132,91	58655,025
[30,45)	37,5	8345,88	312970,5
[45,55)	50	4438,96	221948
[55,65)	60	2343,11	140586,6
n.j		18456,37	758668,08

Sueldo	c_i	n_i	$c_i \cdot n_i$
[0, 6)	3	1597,37	4792,11
[6,10)	8	4168,89	33351,12
[10,12)	11	3878,42	42662,62
[12,16)	14	4439,03	62146,42
[16,21)	18,5	2510,89	46451,47
[21,30)	25,5	1433,18	36546,09
[30,40)	35	428,59	15000,65
		18456,37	240950,5

Y calculamos las medias:

$$\bar{x} = \frac{758668{,}08}{18456{,}37} = 41{,}12 \text{ años} \qquad \bar{y} = \frac{240950{,}5}{18456{,}37} = 13{,}06 \text{ miles €}$$

Hay que notar que, aunque el ejercicio no lo pide explícitamente, es conveniente calcular la desviación típica, para conocer una medida de dispersión de los datos.

b) Construye la tabla de frecuencia de la edad de aquellas personas ocupadas que tienen un sueldo de 1200 a 1600 euros. Calcula también la edad media de las personas que cobran entre 1200 y 1600 euros.

Construimos la tabla de frecuencias de la variable Edad condicionada a que la variable Sueldo está comprendido entre 1200 y 1600 euros. Es decir, X / (Y = [1200,1600)). Así pues:

X/(Y = [1200,1600))	c_i	n_i	$c_i \cdot n_i$
[16,25)	20,5	126,47	2592,635
[25,30)	27,5	430,16	11829,4
[30,45)	37,5	2190,02	82125,75
[45,55)	50	1123,69	56184,5
[55,65)	60	568,69	34121,4
n.j		4439,03	186853,69

En consecuencia, la media será:

$$\overline{X/(Y=[12,16)} = \frac{186853,69}{4439,03} = 42,1 \text{ años}$$

c) Construye la tabla de frecuencia del sueldo de aquellas personas ocupadas que tienen 30 años o más. ¿Qué sueldo medio cobran?

Nótese que la variable que condiciona, en este caso la edad, incluye más de un intervalo. Entonces hay que agrupar las frecuencias conjuntas adecuadamente:

EDAD	*[0, 6)*	*[6,10)*	*[10,12)*	*[12,16)*	*[16,21)*	*[21,30)*	*[30,40)*
[30,45)	566,18	1777,07	*1671,91*	2190,02	1248,87	*736,77*	155,06
[45,55)	323,65	797,11	*881,81*	1123,69	724,93	*448,78*	138,99
[55,65)	185,2	430,59	*503,77*	568,69	306,2	*225,13*	123,53
n.j	**1075,03**	**3004,77**	**3057,49**	**3882,4**	**2280**	**1410,68**	**417,58**

En consecuencia, la variable Sueldo condicionada a la edad de 30 años o más será:

Y/(X ≥ 30)	c_i	n_i	$c_i \cdot n_i$
[0, 6)	3	1075,03	3225,09
[6,10)	8	3004,77	24038,16
[10,12)	11	3057,49	33632,39
[12,16)	14	3882,4	54353,6
[16,21)	18,5	2280	42180
[21,30)	25,5	1410,68	35972,34
[30,40)	35	417,58	14615,3
		15127,95	208016,88

Y, por lo tanto, la media aritmética será:

$$\overline{Y/(X>30} = \frac{208016,88}{15127,95} = 13,751 \text{ miles €}$$

d) ¿Qué porcentaje de personas ocupadas tienen 45 años o más y cobran 1600 € o más?

Hay que buscar el número de ocupados que cumpla las dos variables. Así, hay –observando la tabla y sumando los números adecuados– 1967,56 miles. Lo que representa un 10,66 %.

e) ¿Qué porcentaje de personas ocupadas tiene 45 años o más o cobra 1600 euros o más?

De la misma manera que en el apartado *f*) del ejercicio 2, hay que contar los ocupados que tienen 45 años o más (celdas con mayor tamaño) o cobran 1600 euros o más (celdas rojas).

SUELDO

EDAD	[0, 6)	[6,10)	[10,12)	[12,16)	[16,21)	[21,30)	*[30,40)*	**ni.**
[16,25)	289,79	490,44	*249,08*	126,47	**38,03**	***1,7***	**0**	**1195,51**
[25,30)	232,55	673,68	*571,85*	430,16	**192,86**	***20,8***	**11,01**	**2132,91**
[30,45)	566,18	1777,07	*1671,91*	2190,02	**1248,9**	***736,77***	**155,06**	**8345,88**
[45,55)	323,65	797,11	*881,81*	1123,69	**724,93**	***448,78***	**138,99**	**4438,96**
[55,65)	185,2	430,59	*503,77*	568,69	**306,2**	***225,13***	**123,53**	**2343,11**
n.j	**1597,37**	**4168,89**	**3878,42**	**4439,03**	**2510,89**	**1433,18**	**428,59**	**18456,37**

Del mismo modo que el ejercicio 2, el número de ocupados que cumplen las dos condiciones son:

(45 años o más) + (1600 euros o más) – (45 años o más y 1600 euros o más)

6782,07 + 4372,66 a 1967,56 = 9187,17.

Por lo tanto, el porcentaje de demandas será un 49,788 %.

f) ¿Qué porcentaje de ocupados tiene menos de 30 años de aquellos que cobran 1200 euros o más?

En primer lugar, hay que saber cuántos ocupados cobran 1200 euros o más. Sumando los valores de la tabla se obtienen 8811,69 miles de personas (recuadro rojo). De este defecto 821,03 (sombreado amarillo) tienen menos de 30 años. Por lo tanto, el porcentaje pedido es el 9,32 %.

SUELDO

EDAD	***[0, 6)***	***[6,10)***	***[10,12)***	***[12,16)***	***[16,21)***	***[21,30)***	***[30,40)***
[16,25)	*289,79*	*490,44*	*249,08*	***126,47***	***38,03***	***1,7***	***0***
[25,30)	*232,55*	*673,68*	*571,85*	***430,16***	***192,86***	***20,8***	***11,01***
[30,45)	*566,18*	*1777,07*	*1671,91*	*2190*	*1248,87*	*736,77*	*155,06*
[45,55)	*323,65*	*797,11*	*881,81*	*1123,7*	*724,93*	*448,78*	*138,99*
[55,65)	*185,2*	*430,59*	*503,77*	*568,69*	*306,2*	*225,13*	*123,53*

Ejercicio 5

La siguiente tabla muestra la población en edad de trabajar analfabeta en las comunidades Valenciana, Madrid, Andalucía y el País Vasco a lo largo de los años 2000-2010 en miles de personas.

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Andalucía	332,28	280,99	342,14	307,41	294,91	247,90	262,25	278,03	294,34	290,37	279,91
Madrid	73,15	58,88	74,06	83,33	71,83	47,35	43,35	40,34	46,68	56,98	72,21
País Vasco	17,69	13,77	14,67	12,56	13,27	9,37	11,34	12,50	13,66	14,23	11,10
C. Valenciana	128,01	96,83	110,81	117,30	114,73	69,46	79,91	79,01	92,45	99,06	85,76

Fuente: INE

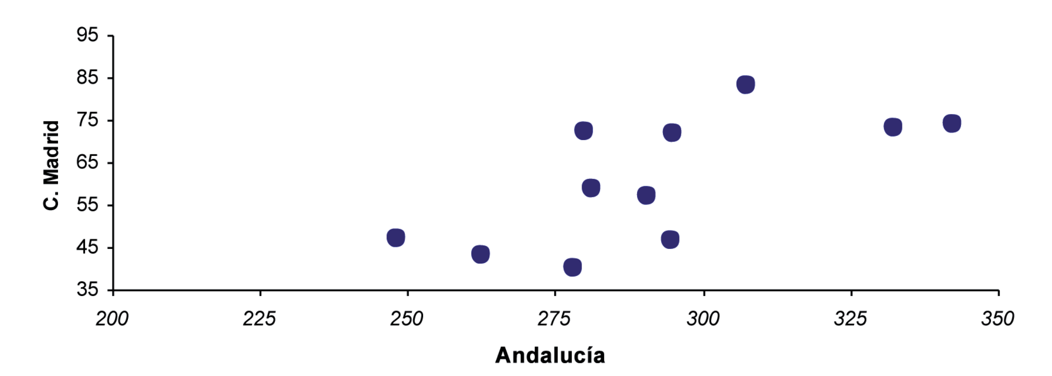
a) Representa la nube de puntos entre las variables: Población en edad de trabajar en Andalucía y Población en edad de trabajar en la Comunidad de Madrid. ¿Qué observas en cuanto a la existencia o no de la relación lineal entre las dos variables?
b) Representa la nube de puntos entre las variables: Población en edad de trabajar en la Comunidad de Madrid y Población en edad de trabajar en el País Vasco. ¿Qué observas en cuanto a la existencia o no de la relación lineal entre las dos variables?
c) Calcula el estadístico adecuado para confirmar las suposiciones que has hecho en los dos apartados anteriores.

Solución

a) Representa la nube de puntos entre las variables: Población en edad de trabajar en Andalucía y Población en edad de trabajar en la Comunidad de Madrid. ¿Qué observas en cuanto a la existencia o no de la relación lineal entre las dos variables?

Una nube de puntos es un gráfico de dos dimensiones en el que se representan los valores de las dos variables. Cada punto de la nube tiene coordenada x (abscisa u horizontal) el valor de una de las variables, y coordenada y (vertical u ordenada) el valor que le corresponde de la otra variable. La forma de este gráfico es el primer paso para saber si dos variables están correlacionadas.

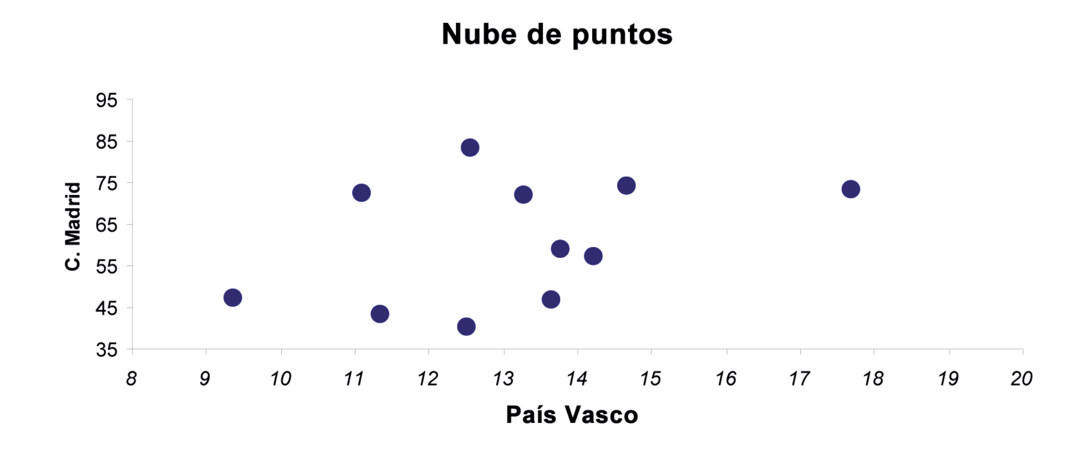
En nuestro caso podemos representar en el eje de las abscisas la población activa y analfabeta en Andalucía, y en el eje de ordenadas las personas analfabetas y activas de la Comunidad de Madrid. Así el gráfico queda:



Con lo que observa visualmente, parece haber una relación lineal positiva entre las dos variables.

b) Representa la nube de puntos entre las variables: Población en edad de trabajar en la Comunidad de Madrid y Población en edad de trabajar en el País Vasco. ¿Qué observas en cuanto a la existencia o no de la relación lineal entre las dos variables?

La estructura será similar al gráfico anterior.



Se observa que también existe una relación lineal positiva, aunque en este caso no parece tan claro porque la nube de puntos es más «ancha». Es decir, parece que los datos no siguen un línea recta creciente con tanta claridad como en el apartado anterior.

c) Calcula el estadístico adecuado para confirmar las suposiciones que has hecho en los dos apartados anteriores.

En los dos apartados anteriores hemos observado a partir del gráfico que existe una relación lineal positiva entre las dos variables. El estadístico que permite contrastar

esta hipótesis es la covarianza. Si este estadístico tiene signo positivo entonces existe relación lineal entre las dos variables y esta es positiva. Si por el contrario, tiene signo negativo entonces también existe relación lineal, pero en este caso es negativa. Si la covarianza es cero, entonces las dos variables no tienen relación lineal.

La expresión de la covarianza será: $S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(x_i - \overline{X})(y_j - \overline{Y})\cdot n_{ij}}{n}$. Sin embargo, para realizar los problemas emplearemos la expresión equivalente:

$$S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} - \overline{X}\cdot\overline{Y}$$

Así pues, hay que calcular las covarianzas de ambos pares de variables.

Andalucía y Comunidad de Madrid

Podemos considerar la variable X = activos y analfabetos en Andalucía y por Y = activos y analfabetos en Madrid.

Como se observa en la fórmula, en primer lugar hay que calcular para cada variable sus medias aritméticas. Haciendo estos cálculos de la misma manera que en la unidad 1 se obtienen los valores: $\overline{X} = 291{,}866$ y $\overline{Y} = 60{,}742$.

En segundo lugar debemos calcular los sumatorios $\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n}$. Como se observa lo que hace falta es multiplicar cada valor de la variable X por su correspondiente de la variable Y y por su frecuencia conjunta. Luego hay que sumar todos estos productos y dividirlos entre el número total de datos.

Como en nuestro caso la frecuencia conjunta de cada dato bivariante es 1, solo hay que hacer los productos de cada valor de una variable por su correspondiente y luego hacer la suma. Así:

X	332,28	280,99	342,14	307,41	294,91	247,9	262,25	278,03	294,34	290,37	279,91
Y	73,15	58,88	74,06	83,33	71,83	47,35	43,35	40,34	46,68	56,98	72,21
$x_i \cdot y_j$	**24306,28**	**16544,69**	**25338,89**	**25616,48**	**21183,39**	**11738,07**	**11368,54**	**11215,73**	**13739,79**	**16545,28**	**20212,30**

Así, sustituyendo la expresión anterior:

$$\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} = \frac{24306{,}28 + 16544{,}69 + 25338{,}89 + 25616{,}48 + \ldots + 16545{,}28 + 20212{,}30}{11}$$

$$= \frac{197809{,}430}{11} = 17982{,}675$$

Y por tanto:

$$S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} - \overline{X}\cdot\overline{Y} = 17982{,}675 - 291{,}866 \cdot 60{,}742 = 332{,}954.$$

Como la covarianza nos queda positiva, también existe relación lineal positiva. Se concluye pues que el gráfico y el estadístico están en concordancia.

País Vasco y Comunidad de Madrid

Hay que hacer exactamente lo mismo, pero ahora considerando X = activos y analfabetos en el País Vasco y para Y = activos y analfabetos en Madrid. Calculado las medias para cada variable: $\overline{X} = 13{,}105$ y $\overline{Y} = 60{,}742$.

Calculamos ahora $\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n}$ tal y como está hecho con anterioridad:

X	17,69	13,77	14,67	12,56	13,27	9,37	11,34	12,5	13,66	14,23	11,1
Y	73,15	58,88	74,06	83,33	71,83	47,35	43,35	40,34	46,68	56,98	72,21
$x_i \cdot y_j$	**1294,024**	**810,778**	**1086,460**	**1046,625**	**953,184**	**443,670**	**491,589**	**504,250**	**637,649**	**810,825**	**801,53**

Así, sustituyendo la expresión anterior:

$$\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} = \frac{1294{,}024 + 810{,}778 + 1086{,}460 + 1046{,}625 + \ldots + 810{,}825 + 801{,}53}{11}$$

$$= \frac{8880{,}594}{11} = 807{,}236$$

Y por lo tanto:

$$S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} - \overline{X}\cdot\overline{Y} = 807{,}236 - 13{,}105 \cdot 60{,}742 = 11{,}212 .$$

Como la covarianza nos queda positiva, también existe relación lineal positiva. Se concluye pues que el gráfico y el estadístico están en concordancia.

Hay que notar, sin embargo, que con la covarianza todavía no podemos saber el grado de esta relación lineal entre ambas variables. Hay que calcular otro estadístico para estudiarlo: el coeficiente de correlación lineal.

Ejercicio 6

Se recolectaron los valores mensuales de los gastos en publicidad de una compañía ferroviaria y el número de pasajeros a lo largo de 15 meses. Los datos los muestra la tabla:

Publicidad (en miles)	10	12	8	17	10	15	10	14	19	10	11	13	16	10	12
Pasajeros (en miles)	15	17	13	23	16	21	14	20	24	17	16	18	23	15	16

a) Calcula el gasto medio y el número medio de pasajeros.
b) Haz la nube de puntos y calcula la covarianza. ¿Es coherente el valor del estadístico con la nube de puntos?
c) Si para los 15 meses posteriores se prevé que la inversión en publicidad de cada mes aumente un 10 % respecto al mismo mes del período anterior, y también se prevé que este hecho provocará un aumento del 8 % en el número de pasajeros cada mes, ¿cuál será la covarianza en este segundo período?

Solución

a) Calcula el gasto medio y el número medio de pasajeros.

Para calcular el gasto medio y el número medio de pasajeros hay que estudiar las distribuciones marginales de las dos variables. Si llamamos X = Gastos en publicidad y Y = Número de pasajeros, entonces las distribuciones marginales son:

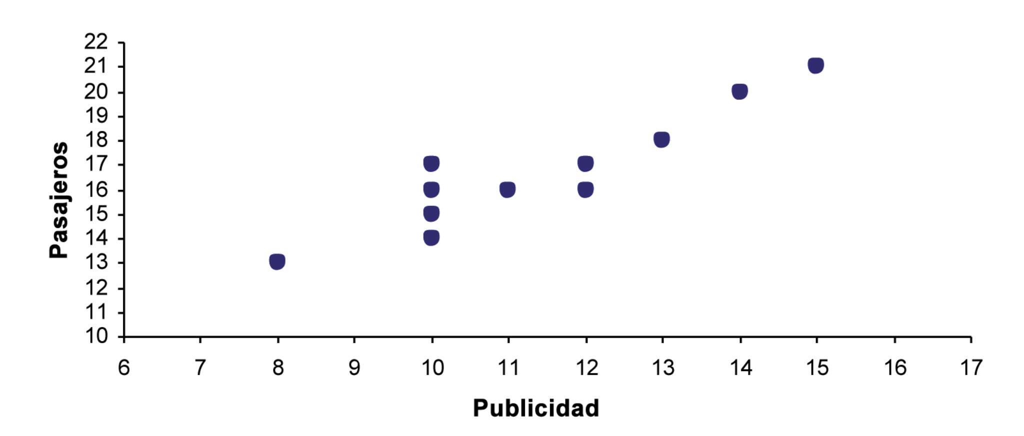
X	ni
8	1
10	5
11	1
12	2
13	1
14	1
15	1
16	1
17	1
19	1
n.j	**15**

Y	**ni**
13	1
14	1
15	2
16	3
17	2
18	1
20	1
21	1
23	2
24	1
n.j	15

Del mismo modo que en la unidad 1 se calculan las medias aritméticas, que en este caso son:

$\overline{X} = 12{,}467$ y $\overline{Y} = 17{,}867$.

b) Haz la nube de puntos y calcula la covarianza. ¿Es coherente el valor del estadístico con la nube de puntos?



Para determinar la covarianza hay que hacer lo mismo que en el ejercicio 5. Como las medias las hemos calculado ya en el apartado anterior, ahora es necesario encontrar otros elementos componentes de la expresión de la covarianza. Así:

X	10	12	8	17	10	15	10	14	19	10	11	13	16	10	12
Y	15	17	13	23	16	21	14	20	24	17	16	18	23	15	16
$x_i \cdot y_j$	**150**	204	104	391	160	315	140	280	456	170	176	234	368	150	192

$\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} = \frac{150+204+.......+150+192}{15} = 233$, por lo tanto la covarianza es:

$$S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} - \overline{X}\cdot\overline{Y} = 233 - 12{,}467 \cdot 17{,}867 = 10{,}252.$$

Como se comprueba, la nube de puntos concuerda con el valor de la varianza.

c) Si para los 15 meses posteriores se prevé que la inversión en publicidad de cada mes aumente un 10 % respecto al mismo mes del período anterior, y también se prevé que este hecho provocará un aumento del 8 % en el número de pasajeros cada mes, ¿cuál será la covarianza en este segundo período?

Se nos pide el valor de la covarianza de las variables Gasto en publicidad y Número de pasajeros en este segundo período; en el que los gastos han aumentado un 10 % y el número de pasajeros un 8 %. Definimos pues estas dos nuevas variables: X' = Gastos en el segundo período y Y' = Número de pasajeros en el segundo período.

Así, según el enunciado X' = 1,1 · X y Y' = 1,08 · Y. Para calcular la covarianza entre X' y Y' únicamente hay que aplicar las propiedades.[8]

Entonces: $S_{X'Y'} = 1{,}1 \cdot 1{,}08\ S_{XY} = 1{,}1 \cdot 1{,}08 \cdot 10{,}252 = 12{,}179$.

8. Si todos los valores de una variable X se multiplican por una constante *a* y todos los valores de la variable Y por una constante *b*, la covarianza queda multiplicada por el producto de las constantes. Es decir: = a · b S_{XY}

Ejercicio 7

Una empresa ha realizado dos tests psicotécnicos a los 9 trabajadores de un departamento como parte del proceso de selección del nuevo director del departamento. La siguiente tabla muestra los resultados obtenidos por los aspirantes:

TEST 1	5	7	6	9	3	1	2	4	6
TEST 2	6	5	8	6	4	2	1	3	7

a) Calcula la covarianza. ¿Existe algún tipo de relación lineal entre las dos variables?

b) Ha habido un error en una pregunta de cada test y el tribunal decide aumentar un 5 % la puntuación de cada participante. Calcula nuevamente la covarianza.

Solución

a) Calcula la covarianza. ¿Existe algún tipo de relación lineal entre las dos variables?

Del mismo modo que en los problemas anteriores, hay que encontrar cada uno de los componentes. Llamamos X = Notas del Test 1 y para Y = Notas del Test 2 y obtenemos las medias de cada variable: $\overline{X} = 4{,}778$ y $\overline{Y} = 4{,}667$.

Buscamos ahora $\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n}$. Para hacerlo, hay que calcular los productos correspondientes y hacer la suma.

X	5	7	6	9	3	1	2	4	6
Y	6	5	8	6	4	2	1	3	7
$x_i \cdot y_j$	30	35	48	54	12	2	2	12	42

Y por lo tanto,

$\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} = \frac{30+35+48+.....+2+12+42}{9} = 26$ y consecuentemente la covarianza es: $S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} - \overline{X}\cdot\overline{Y} = 26 - 4{,}778 \cdot 4{,}667 = 3{,}701$

b) Ha habido un error en una pregunta de cada test y el tribunal decide aumentar un 5 % la puntuación de cada participante. Calcula nuevamente la covarianza.

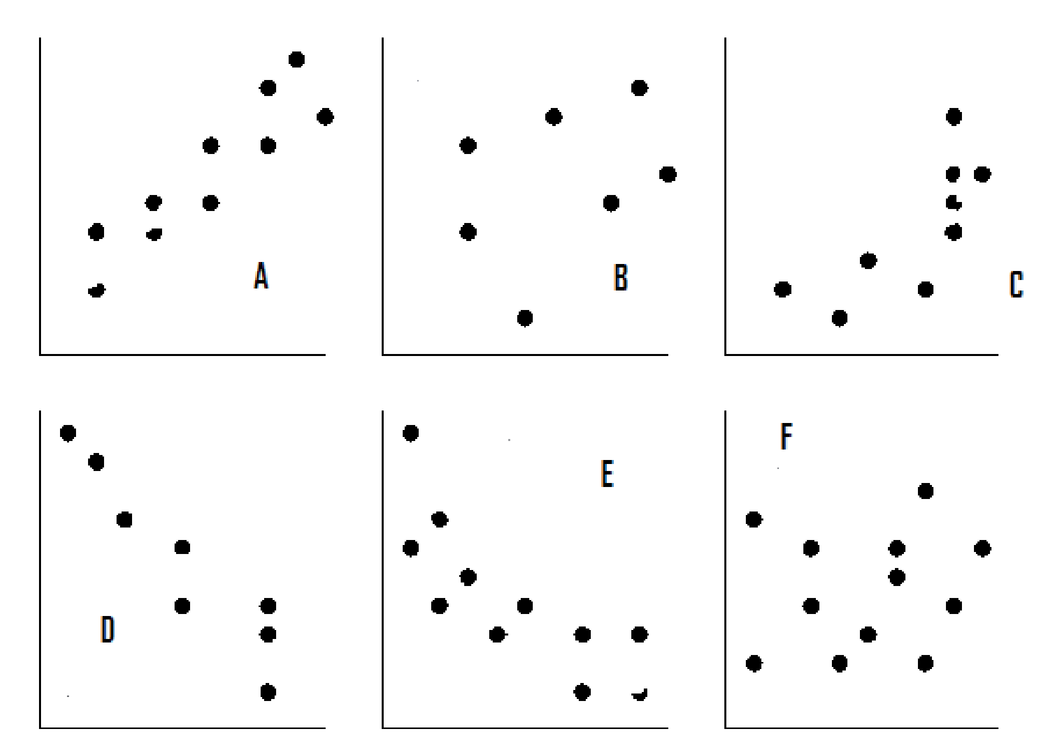
De la misma manera que en el apartado *c*) del ejercicio 6, hay que definir unas nuevas variables y aplicar la misma propiedad. Definimos pues estas dos nuevas variables: X' = Nota del test 1 tras el aumento y Y' = Nota del test 2 después del aumento.

Según el enunciado X' = 1,05 · X y Y' = 1,058 · Y. Para calcular la covarianza entre X' y Y' únicamente hay que aplicar las propiedades.

Entonces: $S_{X'Y'} = 1{,}1 \cdot 1{,}08\ S_{XY}$ = 1,1 · 1,08 · 3,701 = 4,397.

Ejercicio 8

Dadas las siguientes nubes de puntos, contesta:



a) Asocia cada nube de puntos con el valor del coeficiente de correlación que le corresponde entre estos: –0,9; 0,4; 0,95: –0,65; 0,1; 0,6. Razona la respuesta.
b) Indica para cada nube de puntos el signo de la covarianza y di cuál es su significado.

Solución

a) Asocia cada nube de puntos con el valor del coeficiente de correlación que le corresponde entre estos: –0,9; 0,4; 0,95: –0,65; 0,1; 0,6. Razona la respuesta.

Como ya se ha comentado en los problemas anteriores de esta unidad, la covarianza permite discernir si dos variables X y Y tienen una relación positiva, negativa o

cero, pero no aporta información del grado de dependencia de una variable respecto a la otra. Además, la covarianza depende de las unidades de medida empleadas para X y Y. Si por ejemplo X se mide en m^3 y Y en mm^3, cada desviación de X aumenta S_{XY} 10^9 veces. Para hacer frente a estas dos dificultades se define el concepto de correlación lineal r_{XY} :

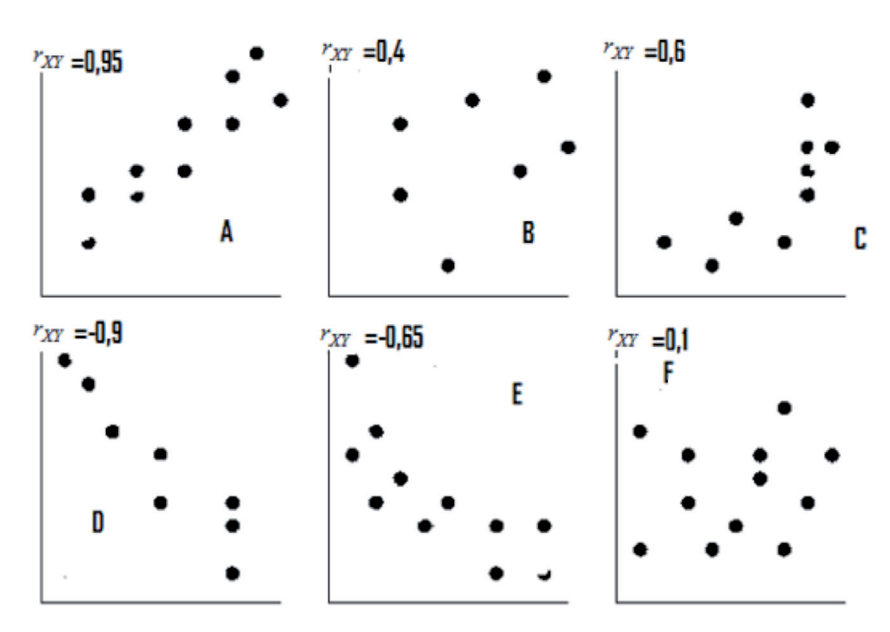
$$r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y}$$ siendo S_X y S_Y las desviaciones típicas de X y Y.

Es evidente que por definición el coeficiente de correlación lineal informa de las mismas cosas que lo hace la covarianza. Además, cumple una propiedad muy importante, está acotado por 1 y por –1. Así pues, se caracteriza por:

- Ser adimensional y siempre estar entre –1 y 1.
- Si hay relación lineal fuerte positiva, $r_{XY} > 0$ y está cerca de 1.
- Si hay relación lineal negativa fuerte, $r_{XY} < 0$ y está cerca de –1.
- Si no hay relación lineal r_{XY} será 0.

Así pues, el coeficiente de correlación será tan próximo a 1 o –1 cuanto la nube de puntos sea más «estrecha». Así pues, a nubes de puntos estrechos le corresponden valores de r_{XY} r_{XY} cercanos a 1 o –1, y por lo tanto los datos se ajustarán bien a una línea recta. Por el contrario, valores r_{XY} próximos al 0, implica que la nube de puntos es más ancha y, en consecuencia, los datos no se ajustaron bien a una línea recta.

En nuestro ejercicio hay que asociar cada nube de puntos con el coeficiente de correlación adecuada. Teniendo en cuenta lo que acabamos de comentar, las asociaciones serán:



b) Indica para cada nube de puntos el signo de la covarianza y di cuál es su significado.

Teniendo presente lo que hemos dicho en el apartado *a*), el signo del coeficiente de correlación es el mismo que el de la covarianza. Por lo tanto, las nubes de los gráficos A, B, C y D tendrán una covarianza positiva (relación lineal positiva) y las nubes E y F tendrán una covarianza negativa (relación lineal negativa). Hay que notar que el gráfico F tiene una correlación de 0,1. Es decir, a pesar de tener signo positivo la relación lineal de ambas variables es muy débil para ser r_{XY} muy cerca de 0.

Ejercicio 9

La siguiente tabla muestra el gasto total promedio, el gasto medio en alimentos y bebidas no alcohólicas y el gasto en vivienda, agua electricidad, gas y otros combustibles en euros, por número de personas que forman la unidad familiar en el año 2009,[9] según datos del INE.

Número de miembros de la familia	*1*	*2*	*3*	*4*	*5*	*6 o más*
Gastos medios totales	18355,25	27755,08	33414,09	38576,14	40699,09	41562,31
Gastos en vivienda, agua, electriciad, gas y otros combustibles	7493,88	8990,72	9205,13	9645,19	10114,49	9272,18

a) ¿Existe una fuerte relación lineal entre el número de miembros que viven en un hogar y el gasto medio total? Razona la respuesta.
b) ¿Y entre el número de miembros que viven en un hogar y el gasto en vivienda, agua, electricidad, gas y otros combustibles? Razona la respuesta.

Solución

a) ¿Existe una fuerte relación lineal entre el número de miembros que viven en un hogar y el gasto medio total? Razona la respuesta.

Ya hemos comentado en el apartado *a*) del ejercicio 8 que para conocer el grado de relación lineal entre dos variables hay que calcular el coeficiente de correlación. La expresión es $r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y}$. Por lo tanto, tenemos que encontrar la covarianza y las desviaciones típicas de cada variable.

Llamamos X = Número de miembros del hogar y Y = Gasto medio total y hacemos los cálculos necesarios. Las medias y las desviaciones típicas las calcularemos tal como hacíamos en la unidad 1. Así: $\overline{X} = 3{,}667$; $S_X = 1{,}972$ i $\overline{Y} = 33393{,}6$; $S_Y = 8214{,}845$.

9. Para realizar el ejercicio, considera 7 miembros en el intervalo *6 o más*.

Tenemos $\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n}$. Para hacerlo es necesario calcular los productos correspondientes y hacer la suma.

X	*1*	*2*	*3*	*4*	*5*	*7*
Y	18355,25	27755,08	33414,09	38576,14	40699,09	41562,31
$x_i \cdot y_j$	18355,25	55510,16	100242,27	154304,56	203495,45	290936,17

Y por lo tanto:

$$\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} = \frac{18355,25+55510,16+.....+203495,45+290936,17}{6} = 137140,643$$

y consecuentemente la covarianza es:

$$S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} - \overline{X}\cdot\overline{Y} = 137140,643 - 3,667\cdot 33363,93 = 14795,112.$$

Así pues, $r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{14795,112}{1,972\cdot 8214,845} = 0,913$, lo que significa que la relación lineal entre las variables Gasto medio total y Miembros que forman un hogar es lo suficientemente fuerte, cuestión perfectamente lógica para otra parte.

b) ¿Y entre el número de miembros que viven en un hogar y el gasto en vivienda, agua, electricidad, gas y otros combustibles? Razona la respuesta.

En este apartado hay que hacer exactamente lo mismo que en el anterior. Llamamos X = Número de miembros del hogar y Y = Gasto en vivienda, agua, electricidad, gas y otros combustibles y hacemos los cálculos necesarios. $\overline{X} = 3,667$; $S_X = 1,972$ y $\overline{Y} = 9120,265$; $S_Y = 812,016$.

Tenemos ahora $\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n}$. Para hacerlo, calculamos los productos correspondientes y la suma.

X	*1*	*2*	*3*	*4*	*5*	*7*
Y	7493,88	8990,72	9205,13	9645,19	10114,49	9272,18
$x_i \cdot y_j$	7493,88	17981,44	27615,39	38580,76	50572,45	64905,26

Y por lo tanto:

$$\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} = \frac{7493,88+17981,44+.....+64905,26}{6} = 34524,8633$$

Y la covarianza será:

$$S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} - \overline{X}\cdot\overline{Y} = 34524{,}863 - 3{,}667\cdot 9120{,}265 = 1080{,}851.$$

De modo que, $r_{XY} = \dfrac{S_{XY}}{S_X \cdot S_Y} = \dfrac{1080{,}851}{1{,}972\cdot 812{,}016} = 0{,}675$, lo que significa que la relación lineal entre las variables Gasto en vivienda, agua, electricidad, gas y otros combustibles y Miembros que forman un hogar es bastante débil.

Ejercicio 10

El director de Recursos Humanos de una empresa ha realizado dos tests psicotécnicos para seleccionar a las personas que deben trabajar en el Departamento de Marketing. Se han presentado 9 personas y los resultados obtenidos en cada uno de los tests han sido los siguientes:

TEST 1	175	181	192	211	235	255	275	286	292
TEST 2	169	185	202	219	240	266	295	329	357

Teniendo en cuenta los resultados de los tests, ¿crees que el director podría haber eliminado uno de los dos tests para decidir los candidatos? Razona la respuesta.

Solución

El director podría haber eliminado una de las dos pruebas siempre y cuando las dos discriminen a las mismas personas. Es decir, si existe una relación casi funcional entre las dos variables. Así pues, para contestar a la pregunta habrá que calcular el coeficiente de correlación lineal y, si este es cercano a 1 o –1, entonces la relación lineal entre las dos variables será fuerte y, por tanto, un test no aportará información adicional y se podría eliminar.

Calculamos, pues, r_{XY} siendo X = notas test 1 y Y = notas test 2.

Hacemos los cálculos necesarios:

$\overline{X} = 233{,}556$; $S_X = 43{,}169$ y $\overline{Y} = 251{,}333$; $S_Y = 61{,}560$.

Buscamos $\displaystyle\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n}$. Para hacerlo hay que calcular los productos correspondientes y hacer la suma.

Y por lo tanto,

$$\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} = \frac{175\cdot 169 + ... + 292\cdot 357}{9} = \frac{551746}{9} = 61305{,}111$$

y consiguientemente la covarianza es:

$$S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} - \overline{X}\cdot\overline{Y} = 61305{,}111 - 233{,}556\cdot 251{,}333 = 2604{,}781.$$

De ahí que, $r_{XY} = \dfrac{S_{XY}}{S_X \cdot S_Y} = \dfrac{2604{,}781}{43{,}169\cdot 61{,}560} = 0{,}980.$

Como el coeficiente r_{XY} es tan cercano a 1, las dos pruebas permiten escoger a las mismas personas, y por tanto una de las dos podría eliminarse.

Ejercicio 11

En una muestra de 150 empresas del sector de servicios se recogen datos sobre el número de trabajadores de la empresa (X) y la facturación (Y) anual en millones de euros. Los resultados se muestran resumidos en los siguientes estadísticos:

$\overline{X}$ = 14 *trabajadores*, $\overline{Y}$ = 100 *millones*, S_X = 2 *trabajadores*; S_Y = 25 *millones*; S_{XY} = 45 *trabajadores* × *millón*

a) Calcula la correlación lineal e interprétalo.
b) Calcula el modelo de regresión lineal que mejor aproxima la facturación en función del número de trabajadores.
c) En función de este ajuste calcula de forma aproximada la cantidad que se espera que facture una empresa con 15 trabajadores. ¿Es fiable esta predicción? Razona la respuesta.
d) Calcula el modelo de regresión lineal que mejor aproxima el número de trabajadores en función de la facturación.
e) En función de este ajuste calcula de forma aproximada el número de trabajadores que se espera que tenga una empresa que facture 105 millones. ¿Es fiable esta predicción? Razona la respuesta.

Solución

a) Calcula la correlación lineal e interprétala.

En este caso hay que únicamente sustituir en la expresión del coeficiente de correlación:

$$r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{45}{25\cdot 2} = 0{,}9.$$

Como está bastante cerca de 1, podemos decir que la relación lineal es bastante fuerte.

b) Calcula el modelo de regresión lineal que mejor aproxima la facturación en función del número de trabajadores

La recta de regresión de una variable Y, llamada explicada o dependiente, respecto a otra X, llamada explicativa o independiente, es la función lineal Y = a*X* + b que mejor se ajusta a los datos empleando el criterio de los mínimos cuadrados. Es decir, por un lado cada valor r_{XY} como está bastante cerca de 1, podemos decir que la relación lineal es bastante fuerte. y_i de la distribución de datos tiene su correspondiente valor x_i por la distribución de datos. Pero además, para todo valor de x_i también se puede calcular su valor por la recta: $y_i^{'} = ax_i + b$. Pues bien, el método de los mínimos cuadrados permite obtener los valores de la ecuación de la recta *a* y *b* que minimizan la suma de los cuadrados de las distancias entre y_i i $y_i^{'}$.

Los valores de *a* y de *b* que se obtienen por el método de los mínimos cuadrados dependen obviamente de los datos. Así, los valores son:

$$a = \overline{y} - \frac{s_{xy}}{s_x^2} \cdot \overline{x} \qquad \text{y} \qquad b = \frac{s_{xy}}{s_x^2}$$

Y, por tanto, la recta de regresión de Y sobre X es:

$$Y = \left(\overline{y} - \frac{S_{XY}}{S_X^2} \cdot \overline{x} \right) + \frac{S_{XY}}{S_X^2} \cdot X$$

Recolocándo los términos, tenemos:

$$Y - \overline{Y} = \frac{S_{XY}}{S_X^2} \cdot (X - \overline{X}).$$

Si se hubiera tomado Y como variable independiente o explicativa, y X como dependiente o explicada, la recta de regresión que se necesita es la que minimiza errores de la X. Se llama recta de regresión de X sobre Y y se calcula fácilmente permutando los puestos de *x* e *y*, obteniéndose:

$$X - \overline{X} = \frac{S_{XY}}{S_Y^2} \cdot (Y - \overline{Y}).$$

En este caso pide la recta que cuenta la facturación en función del número de trabajadores. Por lo tanto, nos pide la recta Y sobre X. Para calcularla, tan solo hay que hacer las sustituciones correspondientes, ya que el ejercicio nos da todos los estadísticos necesarios.

Sustituyendo:

$y-\overline{Y}=\frac{S_{XY}}{S_X^2}\cdot(x-\overline{X})\rightarrow y-100=\frac{45}{4}\cdot(x-14)\rightarrow$ aislando la variable y mediante matemáticas elementales, obtenemos: Y = 11,25 X – 57,5.

Es decir:

La facturación = 11,25 · Número de trabajadores – 57,5.

a) En función de este ajuste calcula de forma aproximada la cantidad que se espera que facture una empresa con 15 trabajadores. ¿Es fiable esta predicción? Razona la respuesta.

Para hacer la predicción únicamente hay que sustituir X por 15, ya que de esta manera obtendremos la estimación de la facturación para una empresa que tuviera 15 trabajadores. Así pues:

La facturación = 11,25 · 15 – 57,5 = 111,25 millones.

Para calcular la fiablilidad hay que emplear el *coeficiente de determinación lineal* (R^2), el cual se puede definir como el porcentaje de varianza de Y que se puede explicar por X, y se le suele llamar calidad o bondad del ajuste porque valora la proximidad de la nube de puntos en la recta de regresión (o dicho con otras palabras, cómo está de ajustada la nube de puntos en la recta de regresión). En las regresiones lineales, este coeficiente tiene una expresión extremadamente simple, ya que coincide con el cuadrado del coeficiente de correlación lineal: $r_{XY}{}^2 = R^2$.

Así, en nuestro caso el coeficiente de determinación será $R^2 = r_{XY}{}^2 = 0{,}9^2 = 0{,}81$ y, por tanto, la fiabilidad es bastante elevada.

b) Calcula el modelo de regresión lineal que mejor aproxima el número de trabajadores en función de la facturación.

En este caso se pide la recta que concreta el número de trabajadores en función de la facturación. Por lo tanto, nos pide la recta X sobre Y. Para calcularla, tan solo hay que hacer las sustituciones correspondientes, ya que el ejercicio nos da todos los estadísticos necesarios.

Sustituyendo:

$X-\overline{X}=\frac{S_{XY}}{S_Y^2}\cdot(Y-\overline{Y})\rightarrow X-14=\frac{45}{625}\cdot(Y-100\rightarrow$ aislando la variable y mediante matemáticas elementales, obtenemos: X = 0,072 Y + 6,8.

Es decir,

Número de trabajadores = 0,072 · la facturación + 6,8.

c) En función de este ajuste calcula de forma aproximada el número de trabajadores que se espera que tenga una empresa que facture 105 millones. ¿Es fiable esta predicción? Razona la respuesta.

Para hacer la predicción únicamente hay que sustituir Y por 105, ya que de esta manera obtendremos la estimación del número de trabajadores para una empresa que tuviera 105 millones de facturación. Así pues:

$$\text{Número de trabajadores} = 0{,}072 \cdot 105 + 6{,}8 = 14{,}36$$

El coeficiente de determinación es $R^2 = r_{XY}^{\ 2} = 0{,}9^2 = 0{,}81$ y, por tanto, la fiabilidad es bastante elevada.

Ejercicio 12

Las dos tablas siguientes muestran el grado medio de satisfacción de los ocupados según el trabajo que realizan por edad y por el nivel de estudios en 2010. Los datos han sido extraídos del Ministerio de Trabajo e Inmigración.

NIVEL ESTUDIOS	***GRADO DE SATISFACCIÓN***
1	7,05
2	7,09
3	7,21
4	7,23
5	7,50
6	7,55

EDAD	***GRADO DE SATISFACCIÓN***
[16,25)	7,33
[25,30)	7,39
[30,45)	7,37
[45,55)	7,30
[55,65)[10]	7,43

Hay que decir que la variable Nivel de estudios ha sido convertida a numérica discreta para ser graduable. Así la equivalencia es: *1 = menos que Primarios; 2 = Primarios; 3 = Secundarios; 4 = Bachillerato; 5 = Formación Profesional y 6 = Universitarios.* Esta conversión se ha hecho a efectos didácticos.

a) Calcula el coeficiente de relación lineal de ambas parejas de variables. ¿En cuál de las dos convendría calcular la recta de regresión?
b) Calcula la recta de regresión del grado de satisfacción en función del nivel de estudios.

Solución

a) Calcula el coeficiente de relación lineal de ambas parejas de variables. ¿En cuál de las dos convendría calcular la recta de regresión?

10. En la tabla original el último intervalo es 55 años o más. Se ha cerrado el intervalo para poder hacer el ejercicio.

Edad y grado de satisfacción

Para calcular el coeficiente de correlación hay que hacer lo mismo que en los ejercicios anteriores. Sin embargo, la variable Edad está agrupada y, por tanto, hay que obtener previamente las marcas de clase.

Llamamos X = Edad y Y = Grado de satisfacción y calculamos los estadísticos necesarios para obtener r_{XY} .

Así, la tabla de la variable X en la que aparecen los intervalos y las clases, y donde se muestran también los productos de los valores de cada variable, así como el sumatorio es:

X	c_i	Y	$x_i \cdot y_j$
[16,25)	20,5	7,33	150,265
[25,30)	27,5	7,39	203,225
[30,45)	37,5	7,37	276,375
[45,55)	50	7,3	365
[55,65)	60	7,43	445,8
			1440,665

De la tabla podemos obtener los estadísticos que se necesitan:

$$\overline{X} = 39{,}1 \; S_X = 14{,}413 \;\text{ i }\; \overline{Y} = 7{,}364 \, S_Y = 0{,}045$$

Tenemos $\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n}$. Para hacerlo hay que calcular los productos correspondientes y hacer la suma.

Por lo tanto:

$$\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} = \frac{20{,}5 \cdot 7{,}33 + ... + 60 \cdot 7{,}43}{5} = \frac{1440{,}665}{5} = 288{,}133$$

y consiguientemente la covarianza es:

$$S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} - \overline{X} \cdot \overline{Y} = 288{,}133 - 39{,}1 \cdot 7{,}364 = 0{,}2006.$$

De modo que $r_{XY} = \dfrac{S_{XY}}{S_X \cdot S_Y} = \dfrac{0{,}2006}{14{,}413 \cdot 0{,}045} = 0{,}309$

Como el coeficient r_{XY} es tan cercano a 0, la relación lineal entre las dos variables es débil.

Nivel de estudios y grado de satisfacción

Llamamos X = Nivel de estudios y Y = Grado de satisfacción y calculamos los estadísticos necesarios para obtener r_{XY}:

$$\overline{X} = 3{,}5 \; S_X = 1{,}708 \text{ y } \overline{Y} = 7{,}272 \; S_Y = 0{,}190$$

Tenemos $\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n}$. Para hacerlo hay que calcular los productos correspondientes y hacer la suma.

Y por lo tanto,

$$\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} = \frac{1 \cdot 7{,}05 + ... + 6 \cdot 7{,}55}{6} = \frac{154{,}58}{6} = 25{,}763$$

y consecuentemente, la covarianza es:

$$S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} - \overline{X} \cdot \overline{Y} = 25{,}763 - 3{,}5 \cdot 7{,}272 = 0{,}311.$$

Por consiguiente, $r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{0{,}311}{1{,}708 \cdot 0{,}190} = 0{,}96.$

Como el coeficiente r_{XY} es tan cercano a 1, la relación lineal entre las dos variables es fuerte y tiene perfecto sentido calcular la recta de regresión.

b) Calcula la recta de regresión del grado de satisfacción en función del nivel de estudios.

Manteniendo la notación del apartado anterior, X = Nivel de estudios y Y = Grado de satisfacción, el ejercicio pide la recta de regresión de Y sobre X. Hay que sustituir los estadísticos calculados en el apartado anterior a la ecuación de la recta:

$y - \overline{Y} = \frac{S_{XY}}{S_X^2} \cdot (x - \overline{X}) \rightarrow y - 7{,}271 = \frac{0{,}311}{2{,}917} \cdot (x - 3{,}5) \rightarrow$ aislando la variable y mediante matemáticas elementales, obtenemos: Y= 0,107 X + 6,898.

Es decir,

La satisfación media en el trabajo = 0,107 · Nivel de estudios + 6,898.

Ejercicio 13

El grado medio de satisfacción medio de los ocupados según el trabajo que realizan por nivel de ingresos y por sexo en el año 2010 se muestra en la tabla siguiente. Los datos han sido extraídos del Ministerio de Trabajo e Inmigración.

NIVEL DE INGRESOS	***GRADO DE SATISFACCIÓN HOMBRES***	***GRADO DE SATISFACCIÓN MUJERES***
[0,600)	6,19	7,253
[600,1000)	6,83	7,234
[1000,1200)	7,28	7,339
[1200,1600)	7,39	7,61
[1600,2100)	7,60	7,768
[2100,3000)	7,82	7,682
[3000,4000)[10]	7,925	7,499

a) Calcula el coeficiente de correlación lineal entre las variables Nivel de ingresos y Grado de satisfacción en los hombres, y entre las variables Nivel de ingresos y Grado de satisfacción en las mujeres. ¿Qué conclusiones se pueden obtener?

b) Calcula la recta de regresión que explique el grado de satisfacción medio en el trabajo de los hombres en función del nivel de ingresos.

Solución

a) Calcula el coeficiente de correlación lineal entre las variables Nivel de ingresos y Grado de satisfacción en los hombres, y entre las variables Nivel de ingresos y Grado de satisfacción en las mujeres. ¿Qué conclusiones se pueden obtener?

Si llamamos por X = Nivel de ingresos, Y = Grado de satisfacción medio en los hombres y Z = Grado de satisfacción medio en las mujeres, se nos pide r_{XY} y r_{XZ}. Para calcularlos hay que hacer lo mismo que en el ejercicio anterior, ya que la variable X está agrupada en intervalos, siendo necesario obtener las marcas de clase. Así:

11. En la tabla original el últímo intervalo es 3000 o más edad. Se ha cerrado el intervalo para poder hacer el ejercicio.

X	c_i	Y	Z
[0,600)	**300**	6,19	7,253
[600,1000)	**800**	6,83	7,234
[1000,1200)	**1100**	7,28	7,339
[1200,1600)	**1400**	7,39	7,61
[1600,2100)	**1850**	7,6	7,768
[2100,3000)	**2550**	7,82	7,682
[3000,4000)	**3000**	7,925	7,499

Nivel de ingresos y grado de satisfación en los hombres

De la tabla podemos obtener los estadísticos que se necesitan:

$$\overline{X} = 1571,429 \; S_X = 889,565 \text{ y } \overline{Y} = 7,291 S_Y = 0,562$$

Ahora buscamos $\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n}$. Para hacerlo hay que calcular los productos correspondientes y hacer la suma.

Y por lo tanto,

$$\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} = \frac{300 \cdot 6,19 + \ldots + 3000 \cdot 7,925}{7} = \frac{83451}{7} = 11921,571$$

y consiguientemente la covarianza es:

$$S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} - \overline{X} \cdot \overline{Y} = 11921,571 - 1571,429 \cdot 7,291 = 464,283.$$

Así que $r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{464,283}{889,565 \cdot 0,562} = 0,93.$

Como el coeficiente r_{XY} es tan cercano a 1, la relación lineal entre las dos variables es fuerte.

Nivel de ingresos y grado de satisfación en las mujeres

En este caso los valores de los estadísticos son: $\overline{X} = 1571{,}429$; $S_X = 889{,}565$

$\overline{Z} = 7{,}484$; $S_z = 0{,}197$ y $\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} = \frac{83146{,}9}{7} = 11878{,}129$

y consiguientemente la covarianza es:

$$S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} - \overline{X} \cdot \overline{Y} = 11878{,}129 - 1571{,}429 \cdot 7{,}484 = 117{,}554.$$

Y por lo tanto, $r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{117{,}554}{889{,}565 \cdot 0{,}197} = 0{,}67.$

Como el coeficiente r_{XY} es menor que el anterior, se puede decir que el grado de satisfacción medio en el trabajo está más relacionado con el sueldo para los hombres que para las mujeres.

b) Calcula la recta de regresión que explique el grado de satisfacción medio en el trabajo de los hombres en función del nivel de ingresos.

Manteniendo la notación del apartado anterior, X = Nivel de ingresos; Y = Grado de satisfacción medio en los hombres, el ejercicio pide la recta de regresión de Y sobre X. Hay que sustituir los estadísticos calculados en el apartado anterior en la ecuación de la recta:

$y - \overline{Y} = \frac{S_{XY}}{S_X^2} \cdot (x - \overline{X}) \rightarrow y - 7{,}291 = \frac{464{,}283}{791325{,}889} \cdot (x - 1571{,}429) \rightarrow$ aislando la variable y mediante matemáticas elementales, obtenemos: Y = 0,000587 X + 6,369.

Es decir:

La satisfación media en el trabajo = 0,000587 · nivel de ingresos + 6,369.

Ejercicio 14

El número total de expedientes de regulación del trabajo a lo largo de los años 2001-2010, según los datos han sido extraídos del Ministerio de Trabajo e Inmigración, son las que se muestran en la tabla.

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Alicante	139	169	224	268	292	180	164	393	939	679
Castellón	49	59	55	76	88	59	58	291	939	777

a) ¿Existe algún tipo de relación lineal entre las variables? ¿Es fuerte esta relación? Razona las respuestas.
b) Calcula la recta de regresión lineal que relaciona el número de expedientes totales en Castellón en función de los de Alicante.

Solución

a) ¿Existe algún tipo de relación lineal entre las variables? ¿Es fuerte esta relación? Razona las respuestas.

Llamamos X = El número total de expedientes de regulación en Castellón Y = El número total de expedientes de regulación en Alicante. Para saber si existe algún tipo de relación lineal entre las variables hay que estudiar la covarianza, y para saber el grado de esta relación lineal, el coeficiente de correlación.

Calculamos, pues, los estadísticos necesarios a partir de la tabla y de la misma manera que hacíamos en la Unidad 1.

$\overline{X} = 344,7 \; S_X = 249,694$, $\overline{Y} = 245,1 S_Y = 316,02$ y

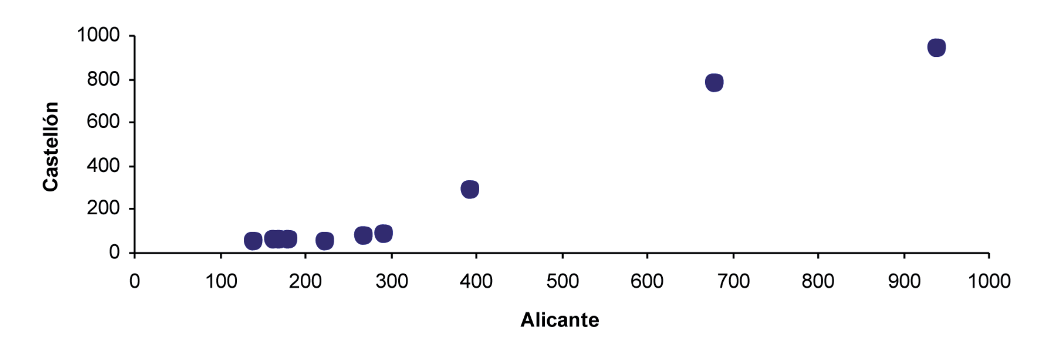
$$\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} = \frac{1618965}{10} = 161896,5$$

y consiguientemente la covarianza es:

$$S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} - \overline{X} \cdot \overline{Y} = 161896,5 - 344,7 \cdot 245,1 = 77410,53.$$

Y por lo tanto, $r_{XY} = \dfrac{S_{XY}}{S_X \cdot S_Y} = \dfrac{77410,53}{249,694 \cdot 316,02} = 0,981.$

Como el coeficiente r_{XY} es muy cercano a 1, existe una correlación lineal muy fuerte entre ambas variables, aunque, como se puede observar, las dos distribuciones marginales tienen mucha dispersión. Es decir, a pesar de que en las dos variables las medias no sean representativas de los datos (valores muy altos de la desviación típica) los datos en conjunto sí que se «amontonan» alrededor de una recta. Este hecho se puede observar en el gráfico de dispersión que aparece a continuación:



b) Calcula la recta de regresión lineal que relaciona el número de expedientes totales en Castellón en función de los de Alicante.

Manteniendo la notación del apartado anterior, de Y sobre X. Hay que sustituir los estadísticos calculados en el apartado anterior a la ecuación de la recta:

$y - \overline{Y} = \frac{S_{XY}}{S_X^2} \cdot (x - \overline{X}) \rightarrow y - 245{,}1 = \frac{77410{,}53}{99868{,}6404} \cdot (x - 249{,}694) \rightarrow$ aislando la variable y mediante matemáticas elementales, obtenemos: Y = 0,775 X + 154,713.

Es decir:

expedientes de Alicante = 0,775 · expedientes de Castellón + 154,713.

Ejercicio 15

La siguiente tabla muestra el número total de hipotecas firmadas, así como la tasa de paro en España en el período 2004-2010, según datos extraídos del INE.

	Hipotecas	***Tasa de paro***
2004	1608497	8,1
2005	1798630	9,2
2006	1896515	8,3
2007	1780627	8,6
2008	1283374	13,9
2009	1082587	18,83
2010	961601	20,05

a) ¿Existe algún tipo de relación lineal entre las variables? ¿Es fuerte esta relación? Razona las respuestas.
b) Calcula la recta de regresión lineal que relaciona el número de hipotecas firmadas en función de la tasa de desempleo.

Solución

a) ¿Existe algún tipo de relación lineal entre las variables? ¿Es fuerte esta relación? Razona las respuestas.

Llamamos Y = Número de hipotecas X = Tasa de paro. Para saber si existe algún tipo de relación lineal entre las variables hay que estudiar la covarianza, y para saber el grado de esta relación lineal, el coeficiente de correlación.

Calculamos, pues, los estadísticos necesarios a partir de la tabla y de la misma manera que hacíamos en la Unidad 1.

$\overline{X} = 12{,}426$; $S_X = 4{,}812$, $\overline{Y} = 1487404{,}429$; $S_y = 347819{,}845$ y

$$\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} = \frac{118134800{,}26}{7} = 16876400{,}037.$$

Y consecuentemente, la covarianza es:

$$S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} - \overline{X}\cdot\overline{Y} = 16876400{,}037 - 12{,}426 \cdot 1487404{,}429 = -1606087{,}405.$$

Y por lo tanto, $r_{XY} = \dfrac{S_{XY}}{S_X \cdot S_Y} = \dfrac{-1606087{,}405}{4{,}812 \cdot 347819{,}845} = 0{,}960.$

Entre estas dos variables existe una relación lineal negativa, ya que la covarianza es menor que cero. Además, como el coeficiente de correlación está muy cerca de –1, la relación lineal es fuerte.

b) Calcula la recta de regresión lineal que relaciona el número de hipotecas firmadas en función de la tasa de desempleo.

Manteniendo la notación del apartado anterior, de Y sobre X. Hay que sustituir los estadísticos calculados en el apartado anterior a la ecuación de la recta:

$y - \overline{Y} = \dfrac{S_{XY}}{S_X^2}\cdot(x - \overline{X}) \rightarrow y - 1487404{,}429 = \dfrac{-1606087{,}405.}{23{,}155}\cdot(x - 12{,}426) \rightarrow$ aislando la variable y mediante matemáticas elementales, obtenemos:

Y= –69362,445 X + 2349302,166, que es la ecuación de la recta que se pide.

Ejercicio 16

La siguiente tabla muestra el número de horas extraordinarias totales en miles (remuneradas y no remuneradas) realizadas en el conjunto de España, así como las tasas de paro desde el primer trimestre de 2008 hasta el último trimeste del año 2010. Los datos han sido extraídos del INE.

Trimestres	*Número total de horas extra*	*Tasa de paro*
2010TIV	5574,9	20,33
2010TIII	5058,9	19,79
2010TII	6002,7	20,09
2010TI	6154,1	20,05
2009TIV	6493,2	18,83
2009TIII	6069	17,93
2009TII	7042	17,92
2008TIV	8398,4	13,91
2008TIII	8813,2	11,33
2008TII	9794,4	10,44
2008TI	10058,1	9,63

a) Halla, en su caso, la recta de regresión que explica el número de horas extras en función de la tasa de desempleo.

b) En el primer trimestre de 2009 la tasa de paro era del 17,36 %. Da una estimación del número de horas extras en este trimestre, así como una medida de su fiabilidad.

Solución

a) Halla, en su caso, la recta de regresión que explica el número de horas extras en función de la tasa de desempleo.

Llamamos X = Tasa de paro y Y = Número de horas extras. El ejercicio pide que calculemos la recta de Y sobre X cuando sea pertinente, es decir, cuando las dos variables estén fuertemente correlada. Por tanto, lo que hay que hacer primero es calcular el coeficiente de correlación entre las dos variables para decidir si es o no pertinente calcular la recta.

P. Juan Verdoy / M. J. Beltrán / M. J. Peris - ISBN: 978-84-15444-38-1 Problemas resueltos de estadística aplicada a las ciencias sociales - UJI - DOI: http://dx.doi.org/10.6035/Sapientia100

Para hallar el coeficiente de correlación hay que calcular algunos estadísticos, tanto de las distribuciones marginales como de la conjunta. así:

$\overline{X} = 16{,}386;\ S_X = 4{,}019;\ \overline{Y} = 7223{,}536;\ S_Y = 1664{,}837$ y

$$\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} = \frac{1230502{,}401}{11} = 111863{,}855$$

y consiguientemente la covarianza es:

$$S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} - \overline{X}\cdot\overline{Y} = 111863{,}855 - 16{,}386 \cdot 7223{,}563 = -6501{,}448.$$

Y por lo tanto, $r_{XY} = \dfrac{S_{XY}}{S_X \cdot S_Y} = \dfrac{-6501{,}448}{4{,}019 \cdot 1664{,}837} = 0{,}972.$

Como el coeficiente de correlación está muy cerca de –1, la relación lineal es fuerte y tiene sentido calcular la recta de regresión Y sobre X.

Calculamos:

$y - \overline{Y} = \dfrac{S_{XY}}{S_X^2}\cdot(x - \overline{X}) \rightarrow y - 7223{,}536 = \dfrac{-6501{,}448.}{16{,}152}\cdot(x - 16{,}386) \rightarrow$ aislando la variable y mediante matemáticas elementales, obtenemos:

Y = –402,516 X + 13819,163, que es la ecuación de la recta de regresión.

b) En el primer trimestre de 2009 la tasa de paro era del 17,36 %. Da una estimación del número de horas extras en este trimestre, así como una medida de su fiabilidad.

Para calcular la estimación debemos sustituir la ecuación de la recta de regresión. Sustituyendo X por 17,36 obtenemos:

Y = –402,516 17,36 + 13819,163 = 6831,485.

Para conocer la fiabilidad de esta predicción es preciso determinar el coeficiente de determinación[12] R^2, el cual es, en las regresiones lineales, el cuadrado del coeficiente de regresión. Por lo tanto, su valor es $R^2 = 0{,}972^2 = 0{,}945$, y la estimación es fiable.

12. Revisar el ejercicio 11.

Ejercicio 17

El número total de expedientes de regulación del trabajo a lo largo de los años 2001-2010 en Cataluña y la Comunidad Valenciana, extraídos del Ministerio de Trabajo e Inmigración son las que se muestran en la tabla, a excepción de los datos del año 2005 que se han omitido.

AÑO	*Cataluña*	*Comunidad Valenciana*
2001	661	465
2002	724	494
2003	608	594
2004	565	619
2006	455	413
2007	470	487
2008	874	1286
2009	3964	3490
2010	3318	2810

Se sabe que el número de expedientes en 2005 en Cataluña fue de 512. Haz una estimación, si conviene, del número de expedientes en la Comunidad Valenciana, así como una medida del ajuste.

Solución

Llamamos X = Número de expedientes en Cataluña y Y = Número de expedientes en la Comunidad Valenciana. Para obtener la estimación de los expedientes en la Comunidad Valenciana a partir de los datos hay que construir la recta de regresión Y sobre X, pero para que este cálculo sea provechoso, es necesario que las dos variables estén fuertemente correlacionadas. Por lo tanto, lo que hay que hacer en primer lugar es calcular el coeficiente de correlación.

Para hallar el coeficiente de correlación hay que calcular algunos estadísticos, tanto de las distribuciones marginales como de la conjunta. Así:

$\overline{X} = 1293{,}2$; $S_X = 1269{,}838$; $\overline{Y} = 1184{,}2$; $S_Y = 1091{,}001$ y la covarianza:

$$S_{XY}=\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n}-\overline{X}\cdot\overline{Y}=2897179{,}67-1293{,}22\cdot 1184{,}22=1365722{,}682.$$

Y por lo tanto, $r_{XY}=\frac{S_{XY}}{S_X \cdot S_Y}=\frac{1365722{,}682}{1269{,}838\cdot 1091{,}001}=0{,}986.$

Como el coeficiente de correlación está muy cerca de 1, la relación lineal es fuerte y tiene sentido calcular la recta de regresión Y sobre X.

Calculamos:

$y-\overline{Y}=\frac{S_{XY}}{S_X^2}\cdot(x-\overline{X})\rightarrow y-1184{,}22=\frac{1365722{,}682.}{1269{,}838^2}\cdot(x-1293{,}22)$ → aislando la variable y mediante matemáticas elementales, obtenemos:

$$Y = 0{,}847\ X + 88{,}907$$

Para estimar el número de expedientes en la Comunidad Valenciana hay que sustituir X por 512, de esta manera obtenemos Y = 0,847 · 512 + 88,907 = 522,571.

Para conocer la fiabilidad de esta predicción es preciso determinar el coeficiente de determinación[13] R^2, el cual es, en las regresiones lineales, el cuadrado del coeficiente de regresión. Por lo tanto, su valor es $R^2 = 0{,}986^2 = 0{,}972$, y la estimación es fiable.

Ejercicio 18

En un museo se desea estudiar la repercusión que tienen las quejas realizadas por los visitantes y los ingresos. Para realizarlo, se observaron las dos variables a lo largo de las últimas diez semanas. Las visitas están expresadas en decenas de asistentes.

Quejas	18	26	30	33	38	39	42	44	46	49
Visitas	107	105,5	105	104,4	104,3	104	103,7	103,4	103,1	103

Si la entrada al museo tiene un coste de 3,6 euros, estima los ingresos del museo si en una semana se hubieran producido 43 quejas.

13. Revisar el ejercicio 11.

Solución

El ejercicio pide que estimemos los ingresos según las quejas. Pero es evidente que los ingresos dependen del número de visitas, por lo tanto, lo que tenemos que averiguar es si existe algún tipo de relación lineal entre el número de quejas y el número de visitas. Si efectivamente esta se produce, entonces podremos encontrar la recta de regresión entre el número de visitas y el de quejas y, con posterioridad, se podrán estimar los ingresos.

Así pues, en primer lugar hay que calcular el coeficiente de correlación entre las dos variables X = Número de quejas y Y = Visitas para saber si están correlacionadas.

Hacemos los cálculos necesarios:

$\overline{X} = 36{,}5$; $S_X = 9{,}211$; $\overline{Y} = 104{,}34$; $S_y = 1{,}166$ y la covarianza:

$$S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} - \overline{X} \cdot \overline{Y} = 3797{,}82 - 36{,}5 \cdot 104{,}34 = -10{,}59.$$

Y por tanto, $r_{XY} = \dfrac{S_{XY}}{S_X \cdot S_Y} = \dfrac{-10{,}59}{9{,}211 \cdot 1{,}166} = -\ 0{,}986.$

Como el coeficiente de correlación está muy cerca a –1, la relación lineal es fuerte y tiene sentido calcular la recta de regresión Y sobre X.

Calculamos:

$y - \overline{Y} = \dfrac{S_{XY}}{S_X^2} \cdot (x - \overline{X}) \rightarrow y - 104{,}34 = \dfrac{-10{,}59.}{9{,}211^2} \cdot (x - 36{,}5) \rightarrow$ aislando la variable y mediante matemáticas elementales, obtenemos:

$$Y = 0{,}125\ X + 108{,}866.$$

Para estimar el número de visitas cuando cuando se producen 43 quejas hay únicamente sustituir X por 43. De esta manera obtenemos Y = 0,125 · 43 + 108,866 = 114,241.

Para conocer la fiabilidad de esta predicción es preciso determinar el coeficiente de determinación[14] R^2, el cual es, en las regresiones lineales, el cuadrado del coeficiente de regresión. Por lo tanto su valor es $R^2 = 0{,}986^2 = 0{,}972$. Por tanto, la estimación es fiable.

Finalmente, los ingresos estimados son 3,6 euros multiplicado por el número de visitas estimadas. Es decir, 3,6 · 1088,66 = 3919,176 euros.

---

14. Revisar el ejercicio 11.

Ejercicio 19

La siguiente tabla muestra el número de personas ocupadas distribuidas atendiendo el sueldo neto de la actividad principal que desarrollan (en centenas de euro) y el nivel de estudios que tenían en 2010, según datos recogidos del Ministerio de Trabajo y de Inmigración. Hay que decir, sin embargo, que la variable nivel de estudios ha sido convertida a numérica discreta para ser graduable. Así, la equivalencia es: *1 = menos que Primarios; 2 = Primarios; 3 = Secundarios; 4 = Bachillerato; 5 = Formación Profesional y 6 = Universitarios.* Esta conversión se ha hecho a efectos didácticos.

	SUELDO						
Nivel de estudios	***[0, 6)***	***[6,10)***	***[10,12)***	***[12,16)***	***[16,21)***	***[21,30)***	***[30, 40)***
1	8,75	21,67	*15,00*	7,93	5,65	*0,74*	1,56
2	293,18	790,92	*601,61*	472,64	92,82	*56,30*	3,77
3	538,08	1551,52	*1226,20*	1098,34	340,72	*74,78*	13,56
4	323,39	670,29	*607,31*	709,62	313,35	*142,00*	53,20
5	303,28	801,87	843,80	982,42	444,90	183,90	50,73
6	164,47	439,26	619,51	1155,75	1230,07	919,40	282,14

a) ¿Están relacionadas linealmente el sueldo y el nivel de estudios?

b) Calcula una estimación del sueldo que cobraría una persona ocupada que tuviera un nivel de estudios equivalente a 4,5 así como su fiablidad.

Solución

a) Están relacionadas linealmente el sueldo y el nivel de estudios?

Llamamos X = Nivel de estudios y Y = Sueldo. Debemos calcular el coeficiente de correlación entre ambas variables para saber si las dos variables están linealmente relacionadas y con qué grado.

Para hacer estos cálculos hay que tener presente que la variable Y está agrupada en intervalos y necesitamos las marcas de clase, y que cada dato bivariante tiene una frecuencia absoluta superior a 1. Este último hecho puede complicar los cálculos, por eso es conveniente seguir un criterio a la hora de hacerlos. Nosotros lo haremos en la tabla de doble entrada:

				Y				
X	[0, 6)	[6,10)	[10,12)	[12,16)	[16,21)	[21,30)	[30, 40)	
Marcas-->	**3**	**8**	**11**	**14**	**18,5**	**25,5**	**35**	***n.j***
1	8,75	21,67	*15*	7,93	5,65	*0,74*	1,56	***61,3***
2	293,18	790,92	*601,61*	472,64	92,82	*56,3*	3,77	***2311,24***
3	538,08	1551,52	*1226,2*	1098,34	340,72	*74,78*	13,56	***4843,2***
4	323,39	670,29	*607,31*	709,62	313,35	*142*	53,2	***2819,16***
5	303,28	801,87	843,8	982,42	444,9	183,9	50,73	***3610,9***
6	164,47	439,26	619,51	1155,75	1230,07	919,4	282,14	***4810,6***
$n_{i\cdot}$	***1631,15***	***4275,53***	***3913,43***	***4426,7***	***2427,51***	***1377,12***	***404,96***	**18456,4**
$\sum x_i \cdot n_{i1}$	***6006,13***	***15584,14***	***15262,12***	***18933,31***	***12071,77***	***7341,58***	***2209,07***	
$y_1 \sum x_i \cdot n_{i1}$	***18018,4***	***124673,12***	***167883,32***	***265066,3***	***223327,75***	***187210,29***	***77317,5***	**1063496,7**

De la manera natural se calculan los estadísticos de las distribuciones marginales: $\overline{X} = 4{,}194$; $S_X = 1{,}412$; $\overline{Y} = 12{,}913$; $S_Y = 6{,}478$. Para calcular la covarianza hay que hacer los productos de cada valor por su correspondiente y luego hacer la suma. Así, el procedimiento que empleamos es el siguiente: la fila con $\sum x_i \cdot n_{i1}$ se obtiene multiplicando cada fila por su frecuencia conjunta, por ejemplo, 6006,13 = 1 · 8,75 + 2 · 293,18+ ... + 6 · 164.47. La fila $y_1 \sum x_i \cdot n_{i1}$ se obtiene multiplicando cada resultado por el valor de la primer columna; por ejemplo, 18018,4 = 6006,13 · 3. Así, $\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} = \frac{1063496{,}7}{18456{,}4}$ = 57,622. Como consecuencia, la covarianza es: $S_{XY} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{x_i y_j \cdot n_{ij}}{n} - \overline{X}\cdot\overline{Y}$ = 57,622 – 4,194 · 12,913 =3,465.

Y por lo tanto, $r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{3{,}465}{1{,}412 \cdot 6{,}478}$ = 0,38, el cual es muy bajo, y por eso la relación es muy débil y no tendría mucho sentido hacer la recta de regresión.

b) Calcula una estimación del sueldo que cobraría una persona ocupada que tuviera un nivel de estudios equivalente a 4,5, así como de su fiablidad.

Como se ha comentado en el apartado anterior, el coeficiente de correlación es muy bajo y, por tanto, carece de sentido hacer la recta de regresión. No obstante, la calcularemos para fines didácticos.

Debemos calcular la recta Y sobre X. Entonces:

$y - \overline{Y} = \frac{S_{XY}}{S_X^2}\cdot(x - \overline{X}) \rightarrow y - 12{,}913 = \frac{3{,}465.}{1{,}412^2}\cdot(x - 4{,}194)$ → aislando la variable y mediante matemáticas elementales, obtenemos:

Y = 1,738 X + 5,622.

Para estimar el sueldo de una persona con un nivel de estudios de 4,5: Y = 1,738 · 4,5 + 5,622 = 13,443 miles de euros.

Para conocer la fiabilidad de esta predicción es preciso determinar el coeficiente de determinación[15] R^2, el cual es, en las regresiones lineales, el cuadrado del coeficiente de regresión. Por lo tanto su valor es $R^2 = 0.38^2 = 0{,}144$, y la estimación es poco fiable.

15. Revisar el ejercicio 11.

UNIDAD 3

Números índice

Introducción teórica

Como elementos introductorios de este capítulo, es conveniente recordar definiciones de conceptos que necesitaremos para alcanzar los objetivos de esta unidad (referencias bibliográficas 7, 11 y 18).

Índice simple

Los índices, calculados a partir de una serie de datos de una magnitud y en un período *t* que denotaremos por $\boldsymbol{x}_{it}$, nos permitirán evaluar, en términos relativos o porcentajes, la evolución de los datos de la serie por períodos. Así:

$$I_{it-1}^{t} = \frac{x_{it}}{x_{it-1}}$$

Índice complejo

Hablaremos de índice complejo cuando queramos estudiar la evolución de una magnitud compleja, porque nos interesa aglutinar en una sola, la diversidad de distintas magnitudes simples. En un ejemplo para analizar la producción de cereales en cierta comunidad (magnitud compleja) necesitamos los datos de maíz, trigo, avena, etc. (que serían las magnitudes simples).

Dentro de los índices complejos, nosotros trabajaremos con índices ponderados. Veremos en los ejercicios las fórmulas de los índices de precios y cantidades de Laspeyres y Paasche, con la intención de analizar las razones que justifican la utilización del índice de Laspeyres de precios en el cálculo oficial del IPC en Europa. Así:

Índice de precios por Laspeyres

Para calcular estos índices, empezaremos por conocer y deducir la fórmula que se empleará. Hemos reducido su cálculo a tres artículos, pero no olvidemos que este cálculo se extiende a la totalidad de artículos representativos del consumo de las familias en un país (Véase ECPF).

$$L_{0\,\text{precios}}^{t} = \frac{\sum_i \frac{p_{it}}{p_{i0}} \cdot p_{i0} \cdot q_{i0}}{\sum_i p_{i0} \cdot q_{i0}} = \frac{\sum_i p_{it} \cdot q_{i0}}{\sum_i p_{i0} \cdot q_{i0}}$$

donde *t* es el año actual y *0* será el año que tomaremos como referencia en la comparación. Si se trata de índices encadenados podríamos decir años *t-1* y *t*.

p_{it} será el precio del artículo y el año *t*
p_{i0} será el precio del artículo y el año *0*
q_{it} será la cantidad del artículo y el año *t*
q_{i0} será la cantidad del artículo y el año *0*

es una media ponderada donde el «peso» de cada artículo, $p_{i0} \cdot q_{i0}$, es el valor del artículo en la «cesta de la compra» del año de referencia y permanecerá constante a lo largo del período, mientras no se cambia la base. Un inconveniente de este método es que si la importancia de los artículos en los hábitos de consumo cambia mucho, estos coeficientes quedan desfasados.

Índice de cantidades por Laspeyres

En este caso se estudia la evolución de las cantidades demandadas y para la ponderación se utilizan los mismos coeficientes de la fórmula anterior $p_{i0} \cdot q_{i0}$:

$$L_{0\,\text{cantidades}}^{t} = \frac{\sum_i \frac{q_{it}}{q_{i0}} \cdot p_{i0} \cdot q_{i0}}{\sum_i p_{i0} \cdot q_{i0}} = \frac{\sum_i q_{it} \cdot p_{i0}}{\sum_i p_{i0} \cdot q_{i0}}$$

Índice de precios por Paasche

$$P_{0\,\text{precios}}^{t} = \frac{\sum_i \frac{p_{it}}{p_{i0}} \cdot p_{i0} \cdot q_{it}}{\sum_i p_{i0} \cdot q_{it}} = \frac{\sum_i p_{it} \cdot q_{it}}{\sum_i p_{i0} \cdot q_{it}}$$

es una media ponderada donde el «peso» de cada artículo, $p_{i0} \cdot q_{it}$, intenta mejorar la propuesta de Laspeyres, evitando en cierto modo el desfase, ya que recoge la importancia del artículo al considerar la cantidad en el período por comparar. Presenta otros inconvenientes.

Índice de cantidades por Paasche

$$P_{\text{cantidades}\ 0}^{\ t} = \frac{\sum_i \frac{q_{it}}{q_{i0}} \cdot q_{i0} \cdot p_{it}}{\sum_i q_{i0} \cdot p_{it}} = \frac{\sum_i p_{it} \cdot q_{it}}{\sum_i q_{i0} \cdot p_{it}}$$

Esta propuesta, como analiza la evolución de las cantidades, considera como coeficiente $p_{i0} \cdot q_{it}$ que indica el «peso» de cada artículo, y el precio del año *t* para actualizar la importancia del artículo.

Índices encadenados

Para calcular el incremento de una serie en un período más largo, podemos utilizar los índices previamente conocidos de los períodos que internamente conforman el período total. Así:

$$I_t^m = I_t^{t+1} \cdot I_{t+1}^{t+2} \cdot I_{t+2}^{t+3} \cdot \ldots \cdot I_{m-2}^{m-1} \cdot I_{m-1}^m$$

Incremento medio

En cierto período largo que incluye varios períodos menores internamente, podemos conocer los incrementos de los períodos menores, uno a uno, consecutivamente y observar entre ellos acusadas diferencias en signos (aumentos y disminuciones) o en magnitud. Para calcular el incremento medio de todos ellos podemos hacer la raíz del producto de estos índices, teniendo en cuenta que el índice de la raíz coincide con el número de índices considerados. Así:

$$I^m_{t\,media} = \sqrt[m-t]{I_t^{t+1} \cdot I_{t+1}^{t+2} \cdot I_{t+2}^{t+3} \cdot \ldots \cdot I_{m-2}^{m-1} \cdot I_{m-1}^m}$$

Canvio de base en el IPC

Por razones que habrá que profundizar en la teoría, en ciertos momentos había que hacer un cambio en el año de referencia del cálculo oficial del IPC y se empezaba a obtener la nueva serie de este índice, comenzado de nuevo con el valor 100. Diremos que se hacía un «cambio de base».

A menudo, como podremos ver en los ejercicios propuestos, es necesario utilizar en un mismo cálculo el valor del IPC de años que corresponden a períodos de bases diferentes y necesitaremos trabajar con todos los valores del IPC referidos a una misma base. Estos datos los podrás encontrar fácilmente en la página web del INE, pero en los ejercicios podremos ver cómo se puede encontrar el enlace técnico que permite unificar la base de las dos series.

Para obtener los valores del IPC del año «y» en base B hemos multiplicado el IPC del año «y» en base A por la fracción $\frac{IPC_B^B}{IPC_A^B}$ el valor de la que se lo que denominaremos «enlace técnico».

Sería interesante profundizar en los mecanismos del cálculo del IPC por el INE y los cambios metodológicos introducidos en los últimos años a partir del 2000 y el proceso de armonización con Europa.

http://www.ine.es/daco/daco43/metoipc06.pdf

Actualización de un valor

Utilizando los valores del IPC, podemos conocer el valor actualizado de una renta, de un alquiler, de un sueldo, de un bien, etc. Tan solo hemos considerado que esta operación la podremos hacer siempre y cuando los períodos iniciales y finales por considerar estén ambos antes o después de enero de 2002. En otro caso, habrá que recurrir a un índice LAU que se puede encontrar en el INE:

$$Valor\ \ actualizado = Valor\ \ inicial \times \frac{IPC\ \ mes\ \ final}{IPC\ \ mes\ \ inicial}$$

Pérdida o gananacia del poder adquisitivo

Se puede decir que nosotros ganamos poder adquisitivo (capacidad de compra de bienes de consumo) si el salario que percibimos este año está por encima de lo que ingresaríamos si nuestro salario hubiera sido incrementado en el mismo porcentaje que aumentan los precios de estos bienes. Podríamos razonar de la misma manera para definir la pérdida de poder adquisitivo cuando nuestro salario queda por debajo de lo que tendríamos si la hubieran incrementado con el mismo porcentaje que los precios.

El incremento de estos precios está reflejado en el IPC que publica el INE cada mes. Nosotros tomaremos la media anual general de este índice que podremos encontrar fácilmente en la web de este organismo.

Ahora bien, como hacemos un análisis en términos relativos y damos el resultado en porcentajes, veamos en la siguiente expresión, cómo el salario concreto del que partimos no es necesario en el estudio de la evolución del poder adquisitivo:

Ganancia o pérdida del poder adquisitivo

$$= \Delta_{\text{poder adquisitivo}} = \frac{salario\ \ nuevo}{salario\ actualizado\ \ según\ IPC} = \frac{salario\ \ anterior \cdot (1 + \Delta_{salarial})}{salario\ \ anterior \cdot (1 + \Delta_{IPC})}$$

$$= \frac{(1 + \Delta_{salarial})}{(1 + \Delta_{IPC})}$$

La pérdida o ganancia del poder adquisitivo, pues, se calcula a partir de la comparación de los incrementos anuales del salario (Δ_{salario}) y del IPC (Δ_{IPC}) paralelamente.

A tal fin, comenzaremos por calcular los índices «encadenados» de los salarios, que nos permitirán averiguar los incrementos salariales anuales, y a partir de los valores del IPC publicados en el INE podremos hacer lo mismo.

Deflactar una serie o pasarla a temas reales

Para estudiar el análisis de una magnitud económica en *términos reales*, es necesario transformar los valores originales en *términos corrientes* mediante los IPC que convengan, para convertir todos los valores de la serie en los equivalentes referidos a un mismo año que denominaremos *año de referencia*. Esta operación es llamada *deflactación* de la serie.

Con esta operación, le hemos «eliminado» a la serie original, el efecto de la inflación y podremos analizar «en términos reales» su evolución como tal magnitud, salvo las influencias de los devenires de la economía general que se reflejan en las variaciones del índice de precios.

Objetivos

Los problemas deben permitir que los alumnos alcancen los objetivos didácticos:

3*a*) Saber calcular los números índices simples de una serie de valores para estudiar la evolución de una magnitud a lo largo del tiempo.
3*b*) Interpretar el valor del índice para conocer el incremento porcentual de la magnitud en el período indicado y viceversa.
3*c*) Saber calcular índices con la misma base de referencia y también índices encadenados.
3*d*) Calcular el incremento total y medio de una serie en cierto período, así como los índices correspondientes, tanto si conocemos los términos de una serie como si conocemos sus incrementos porcentuales por períodos.
3*e*) Conocer y calcular el enlace para cambiar de base los índices.
3*f*) Conocer las fórmulas de Laspeyres y Paasche como índices complejos.
3*g*) Actualizar el valor de un bien utilizando los valores del IPC.
3*h*) Deflactar una serie utilizando el IPC.
3*i*) Conocer el concepto de términos monetarios nominales (moneda corriente) y términos reales (moneda constante) en una serie económica para evaluar su evolución.
3*j*) Hacer previsiones de los valores de una serie para datos inmediatos.
3*k*) Saber calcular las variaciones del poder adquisitivo de un salario, en función de las variaciones del salario y del IPC.

La tabla siguiente nos muestra cómo están distribuidos los objetivos según los ejercicios:

Objetivos **Ejercicio**	***3a***	***3b***	***3c***	***3d***	***3e***	***3f***	***3g***	***3h***	***3i***	***3j***	***3k***
1	**x**	**x**	**x**	**x**						x	
2	x	x	x	x						x	
3	x	x	x	x						x	
4					x						
5						x					
6							x				
7				x				x	x	x	
8											x
9											x
10	x	x	x	x	x			x	x	x	

Enunciados

3*a*) Saber calcular los números índices simples de una serie de valores para estudiar la evolución de una magnitud a lo largo del tiempo.
3*b*) Interpretar el valor del índice para conocer el incremento porcentual de la magnitud en el período indicado y viceversa.
3*c*) Saber calcular índices con la misma base de referencia y también índices encadenados.
3*d*) Calcular el incremento total y medio de una serie en cierto período, así como los índices correspondientes, tanto si conocemos los términos de una serie como si conocemos sus incrementos porcentuales por períodos.
3*j*) Hacer previsiones de los valores de una serie para datos inmediatos.

Ejercicio 1

A continuación presentamos el volumen total de alumnos matriculados en la Universitat Jaume I en los últimos años.

	Número total de alumnos matriculados
Curso 2005/2006	12676
Curso 2006/2007	12928
Curso 2007/2008	13159
Curso 2008/2009	13210
Curso 2009/2010	13904
Curso 2010/2011	14702

a) Calcula los índices para cada año, tomando como año de referencia el 2005 (hará referencia al curso 2005-2006). Interpreta el resultado.
b) Calcula los índices encadenados de esta serie. Interpreta los resultados.
c) Calcula el incremento total e incremento medio anual de este período, a partir de las cantidades originales y de los índices encadenados.
d) Haz previsiones para la matrícula del curso 2011/2012 y 2012/2013 si consideramos que no habrán cambios significativos en su comportamiento.

3*a*) Saber calcular los números índices simples de una serie de valores para estudiar la evolución de una magnitud a lo largo del tiempo.
3*b*) Interpretar el valor del índice para conocer el incremento porcentual de la magnitud en el período indicado y viceversa.
3*c*) Saber calcular índices con la misma base de referencia y también índices encadenados.
3*d*) Calcular el incremento total y medio de una serie en cierto período, así como los índices correspondientes, tanto si conocemos los términos de una serie como si conocemos sus incrementos porcentuales por períodos.
3*j*) Hacer previsiones de los valores de una serie para datos inmediatos.

Ejercicio 2

En la siguiente tabla se muestran los datos del INI que hacen referencia al total de visitantes a los parques nacionales de España, en los años que hacemos referencia.

Naturaleza y biodiversidad	
Zonas protegidas	
Número de visitantes por nacionalidades y período	
Unidades: número de personas	
	Total
2000	10252799
2001	10002517
2002	9661493
2003	10296382
2004	11134880
2005	10743480
2006	10979470
2007	10864738
2008	10222818
2009	9952606
Fuente: Ministerio de Medio Ambiente y Medio Rural y Marino. Red de parques Naturales. Copyright INE 2011	

a) Calcula los índices para cada año, tomando como año de referencia el 2000 e interpreta el resultado.
b) Calcula los índices encadenados de esta serie. Interpreta los resultados.
c) Calcula el incremento total e incremento medio anual de este período, a partir de las cantidades originales y de los índices encadenados.
d) Haz previsiones del número de visitantes de los parques considerados para los años 2010, 2011 y 2012, si consideramos que no hubiera cambios significativos en el comportamiento de la afluencia.

Fuente: INE

3*a*) Saber calcular los números índices simples de una serie de valores para estudiar la evolución de una magnitud a lo largo del tiempo.
3*b*) Interpretar el valor del índice para conocer el incremento porcentual de la magnitud en el período indicado y viceversa.
3*c*) Saber calcular índices con la misma base de referencia y también índices encadenados.
3*d*) Calcular el incremento total y medio de una serie en cierto período, así como los índices correspondientes, tanto si conocemos los términos de una serie como si conocemos sus incrementos porcentuales por períodos.
3*j*) Hacer previsiones de los valores de una serie para datos inmediatos.

Ejercicio 3

A continuación presentamos las variaciones porcentuales del volumen de ventas de cierta superficie comercial, en los últimos años.

Año	Variaciones del volumen de ventas (%)
2006	–3,13
2007	–2,15
2008	+2,12
2009	+3,15
2010	+4,12
2011	+4,31

a) Calcula los índices de las ventas de cada año, tomando como referencia el año 2005 y los índices encadenados.
b) Calcula la variación o incremento medio anual y total de las ventas en este período.
c) Estima las ventas de los dos años siguientes si suponemos que no hay cambios significativos en el comportamiento de las ventas en estos años.

3*e*) Conocer y calcular el enlace para cambiar de base los índices.

Ejercicio 4

A continuación presentamos los valores del índice de precios al consumo, IPC, que podemos consultar en la página del INE, y que hace referencia a los datos en base a 2001 y 2006.

Por razones que habrá que estudiar en la teoría, en ciertos momentos hay que hacer un cambio en el año de referencia y se empieza a obtener la nueva serie del IPC, comenzado de nuevo con el valor 100. Diremos que ha habido un «cambio de base».

A menudo, como podrás ver en ejercicios posteriores, hay que utilizar en un mismo cálculo el valor del IPC de años que corresponden a períodos de bases diferentes y necesitaremos trabajar con todos los valores del IPC referidos a una misma base. Estos datos los podrás encontrar fácilmente en la página web del INE, pero en este ejercicio vamos a ver cómo se calculan los valores de las casillas que están sombreadas en gris.

En primer lugar, presentamos la tabla de los valores del IPC desde el año 2002 al 2006 en base 2001.

Índice de Precios al Consumo	
Medias anuales. Base 2001	
Nacional por general y Grupos COICOP	
Unidades: Índices y tasas	
	General
	Media anual
2006	117,624
2005	113,63
2004	109,927
2003	106,684
2002	103,538

Y a continuación, los datos de los valores del IPC desde el año 2006 al 2010 en base 2006, aunque están añadidos los valores de las casillas grises que corresponden a los valores obtenidos «a posteriori» para facilitar los trabajos de cálculo referidos a períodos de diferentes bases.

Índice de Precios al Consumo	
Medias anuales. Base 2006	
Índices nacionales: general y de grupos COICOP	
Unidades: Base 2006 = 100	
	General
	Media anual
2010	108,588
2009	106,668
2008	106,976
2007	102,787
2006	100
2005	96,604
2004	93,456
2003	90,699
2002	88,024

Explica cómo se han obtenido los datos de las casillas sombreadas en gris, averiguando el valor del enlace.

Fuente: INE

3*f*) Conocer las fórmulas de Laspeyres y Paasche como índice complejos.

Ejercicio 5

Calcular los índices de precios y cantidades de los artículos A, B y C mediante la fórmula de Laspeyres y Paasche, de los años 2008, 2009 y 2010 en función del año 2008, utilizando los datos de las siguientes tablas donde están indicadas las cantidades q_i y precios p_i que hay que conocer.

	2008		2009		2010	
	Precio p_i	Cantidad q_i	Precio p_i	Cantidad q_i	Precio p_i	Cantidad q_i
Art. A	12	100	14	112	15	115
Art. B	10	50	8	65	7	72
Art. C	5	20	10	10	15	5

3*g*) Actualizar el valor de un bien, utilizando los valores del IPC.

Ejercicio 6

Supongamos que compramos una vivienda por 16.125.000 ptas. en diciembre de 1998 y la hemos vendido en diciembre de 2006 por un valor de 240.000 euros. Averigua el porcentaje de beneficios o pérdidas que hemos tenido en la operación.

Nota: Para realizar las operaciones consultaremos los valores del IPC que necesitamos en la página web del INE. www.ine.es (sería interesante calcular este incremento con el IPC general y con el IPC del grupo de la vivienda).

El cambio de moneda que consideraremos se 1 € = 166,386 ptas.

3*h*) Deflactar una serie utilizando el IPC.
3*i*) Conocer el concepto de términos monetarios nominales (moneda corriente) y términos reales (moneda constante) en una serie económica para evaluar su evolución.
3*j*) Hacer previsiones de los valores de una serie para datos inmediatos.
3*k*) Saber calcular las variaciones del poder adquisitivo de un salario, en función de las variaciones del salario y del IPC.

Ejercicio 7

En la siguiente tabla mostramos los datos de los impuestos municipales de cierta vivienda en los últimos años.

Año	Importe impuesto municipal (términos nominales)
2006	503,24
2007	515,65
2008	536,73
2009	578,84
2010	584,42

Para analizar su evolución:

a) Deflactar la serie, convirtiéndola en monedas constantes del año 2006.
b) Calcula los índices que nos permitirán estudiar su evolución año por año, en términos reales o monedas constantes del año 2006. Interpreta los resultados.
c) Calcula el incremento total e incremento medio en el período en términos reales.
d) Suponiendo que los impuestos sigan este comportamiento, averigua el valor, en términos nominales o monedas corrientes para los años 2011, 2012 y 2013.

Nota: Para resolver este ejercicio utilizaremos los valores de la media anual del IPC general que necesitamos, obteniéndose de la página web del INE. www.ine.es

3*k*) Saber calcular las variaciones del poder adquisitivo de un salario, en función de las variaciones del salario y del IPC.

Ejercicio 8

En la tabla siguiente se indica el valor de la nómina mensual de un trabajador en los últimos años.

Estudia la pérdida o ganancia de su poder adquisitivo para cada año y de todo el período global, considerando los valores de la media anual del IPC general que puedes encontrar en la página del INE.

Año	Nómina mensual (€)
2007	2034,75
2008	2062,13
2009	2218,61
2010	2253,67
2011	2181,75

3*k*) Saber calcular las variaciones del poder adquisitivo de un salario, en función de las variaciones del salario y del IPC.

Ejercicio 9

En las tablas siguientes se presentan los valores del IPC y el incremento salarial de un trabajador en los años que se indica en cierta comunidad.

Años	IPC
2008	115,1
2009	119,2
2010	121,6
2011	123,8

Años	Incremento salarial anual (%) Anual IPC
2008	
2009	1,8
2010	2,7
2011	1,7

a) Calcula el incremento medio y total del salario en el período 2008-2011.
b) Calcula el incremento anual, medio y total del IPC en el período 2008-2011.
c) Si suponemos que las condiciones económicas de la comunidad no varían, realiza una previsión del valor del IPC para el año 2013.
d) Estudia para cada año y para el período total la pérdida o ganancia del poder adquisitivo y realiza una interpretación de los datos obtenidos.

3*a*) Saber calcular los números índices simples de una serie de valores para estudiar la evolución de una magnitud a lo largo del tiempo.
3*b*) Interpretar el valor del índice para conocer el incremento porcentual de la magnitud en el período indicado y viceversa.
3*c*) Saber calcular índices con la misma base de referencia y también índices encadenados.
3*d*) Calcular el incremento total y medio de una serie en cierto período, así como los índices correspondientes, tanto si conocemos los términos de una serie como si conocemos sus incrementos porcentuales por períodos.
3*e*) Conocer y calcular el enlace para cambiar de base los índices.
3*h*) Deflactar una serie utilizando el IPC.
3*i*) Conocer el concepto de términos monetarios nominales (moneda corriente) y términos reales (moneda constante) en una serie económica para evaluar su evolución.
3*j*) Hacer previsiones de los valores de una serie para datos inmediatos.

Ejercicio 10

Para hacer un estudio de la evolución del precio de cierto modelo de ordenador en términos reales, disponemos de los datos que presentamos en la tabla siguiente:

a) Calcula el incremento anual, medio y total del precio del ordenador en términos reales.
b) Si seguimos esta evolución, estima el precio que podría tener el ordenador en 2008.

c)

	2000	2001	2002	2003	2004
IPC base 1992	131	135	139		
IPC base 2002			103	106	109
	1300	1275	1250	1100	950

Nota: Debemos recurrir a períodos y valores muy antiguos o imaginados para trabajar el objetivo del cambio de base del IPC, debido a que con la nueva metodología del cálculo del IPC por el INE esta circunstancia se ha superado, pero es importante que el alumno conozca este contenido para advertir la necesidad de no trabajar en series de IPC no adecuadas en un mismo ejercicio.

Ayudas

En este apartado se presentarán las ayudas para emplear en caso de ser necesario a la hora de realizar los ejercicios y problemas. Es conveniente no hacer un abuso excesivo de estas ayudas, es decir, antes de emplearlas hay que pensar el problema al menos durante unos 10-15 minutos. Después se consultará la ayuda de tipo 1 y se intentará resolver el ejercicio con esta ayuda. Si no es posible resolverlo, entonces se consultará la ayuda de tipo 2; y en último término la solución.

Ayudas Tipo 1

Ejercicio 1

En el apartado *a*) hay que calcular los índices en base 2005, comparando de manera porcentual cada valor de la lista con 12.676.

En el apartado *b*) hay que hacer lo mismo que en el apartado anterior pero comparando cada valor de la lista con la anterior, utilizando e interpretando los índices correspondientes.

En el apartado *c*) hay que comparar el primer y último valor de la lista para el incremento total y calcular después la raíz correspondiente para calcular el incremento medio anual.

Hay que decidir el índice de esta raíz, en función del número de años que consideramos en el período.

d) Para hacer previsiones de la matrícula para los cursos venideros, utilizaremos el incremento medio anual que hemos obtenido en el apartado anterior.

Ejercicio 2

Todos los apartados de este ejercicio son iguales que los del ejercicio anterior excepto en un aspecto: el primer ejercicio, los valores de la magnitud siempre crecían, mientras que en este segundo ejercicio algunos valores aumentan y en otros disminuyen a lo largo del período considerado.

Ejercicio 3

En este ejercicio, a diferencia de los dos anteriores, no conocen los valores de la magnitud que queremos estudiar. En el enunciado nos dan los porcentajes de crecimiento o decrecimiento directamente. Es necesario que convirtamos esta información directamente en índice. Por ejemplo, puedes rellenar la siguiente tabla para hacer el apartado *a*) y resolver el resto de apartados como lo has aprendido los ejercicios 1 y 2.

Año	Variaciones del volumen de ventas (%)	Índices Encadenados
2006	-3,13	$I_{2005}^{2006} = 0{,}9687$
2007	*-2,15*	$I_{2006}^{2007} =$
2008	*+2,12*	$I_{2007}^{2008} = 1{,}0212$
2009	*+3,15*	$I_{2008}^{2009} =$
2010	*+4,12*	$I_{2009}^{2010} =$
2011	*+4,31*	$I_{2010}^{2011} =$

Es interesante aprender a calcular los índices en base 2005 a partir de los índices encadenados de la tabla anterior. Por ejemplo:

$$I_{2005}^{2007} = I_{2005}^{2006} \cdot I_{2006}^{2007} = 0{,}9687 \cdot 0{,}9785 = 0{,}9479$$

Y así sucesivamente...

$$I_{2005}^{2008} = I_{2005}^{2006} \cdot I_{2006}^{2007} \cdot I_{2007}^{2008} = \ldots$$

$$I_{2005}^{2009} = I_{2005}^{2006} \cdot I_{2006}^{2007} \cdot I_{2007}^{2008} \cdot I_{2008}^{2009} = \ldots$$

$$I_{2005}^{2010} = I_{2005}^{2006} \cdot I_{2006}^{2007} \cdot I_{2007}^{2008} \cdot I_{2008}^{2009} \cdot I_{2009}^{2010} = \ldots$$

$$I_{2005}^{2011} = I_{2005}^{2006} \cdot I_{2006}^{2007} \cdot I_{2007}^{2008} \cdot I_{2008}^{2009} \cdot I_{2009}^{2010} \cdot I_{2010}^{2011} = \ldots$$

En el apartado *b*) nos piden el incremento total, que ya hemos encontrado en la última línea, y el incremento medio anual mediante el cálculo de la raíz sexta del índice que acabamos de nombrar.

En el apartado *c*) hay que hacer una reflexión. Explicar por qué no es posible hacer la previsión que nos piden.

Ejercicio 4

En este ejercicio nos piden un estudio de cómo se hace *un cambio de base* en la secuencia de valores del IPC. Recordar que este cálculo es necesario hacerlo siempre que necesitamos utilizar en una misma operación valores del IPC que corresponden a años que se han calculado con diferentes bases. También hablaremos del *enlace técnico* que permite hacer directamente esta transformación.

Hay que comprender que tan solo se plantea una proporcionalidad (regla de tres) que se mantiene entre la secuencia de valores de la serie de una base y la serie obtenida con la base nueva. También hay que considerar que en el año que se hace el cambio de base, se calcula el IPC con la vieja y nueva base y se plantea la equivalencia.

Veamos el planteamiento del primer año que hay que resolver:

$$IPC_{2001}^{2002} = 103{,}538 \rightarrow IPC_{2006}^{2002} = \boldsymbol{x}$$

$$IPC_{2001}^{2006} = \boldsymbol{117{,}624} \rightarrow IPC_{2006}^{2006} = \boldsymbol{100}$$

Los datos marcados en rojo serán los términos necesarios para definir el enlace técnico.

Ejercicio 5

Para resolver este ejercicio es necesario conocer las fórmulas que hay que emplear, y que puedes consultar en un manual de teoría sobre los índices de precios y cantidades.

Índice de precios de Laspeyres

$$L_{precios_0}^{\ t} == \frac{\sum_i p_{it} \cdot q_{i0}}{\sum_i p_{i0} \cdot q_{i0}}$$

Índice de cualidades de Laspeyres

$$L_{cantidades_0}^{\ t} = \frac{\sum_i q_{it} \cdot p_{i0}}{\sum_i p_{i0} \cdot q_{i0}}$$

Índices de precios de Paasche

$$P_{preus\,0}^{\ t} = \frac{\sum_i p_{it} \cdot q_{it}}{\sum_i p_{i0} \cdot q_{it}}$$

Índice de cantidades de Paasche

$$P_{cantidades_0^t} = \frac{\sum_i p_{it} \cdot q_{it}}{\sum_i q_{i0} \cdot p_{it}}$$

donde *t* es el año actual y *0* será el año que tomaremos como referencia en la comparación. Si se trata de índices encadenados podríamos decir años *t-1* y *t*.

p_{it} será el precio del artículo y el año t
p_{i0} será el precio del artículo y el año 0
q_{it} será la cantidad del artículo y el año t
q_{i0} será la cantidad del artículo y el año 0

Ejercicio 6

En este ejercicio en primer lugar, es necesario actualizar la moneda. Nosotros lo hemos resuelto aplicando el cambio y convirtiendo la cantidad de compra en euros.

Para actualizar el valor, es necesario conocer el valor del IPC del momento de compra y venta en la misma base y operar convenientemente.

Ejercicio 7

En este ejercicio aprenderemos a estudiar el análisis de una magnitud económica en términos reales.

En el apartado *a*) hay que transformar los valores originales mediante el IPC para convertir todos los valores en los equivalentes del año 2006. Esta operación se llama *deflactación* de la serie.

En el apartado *b*) calcularemos los índices encadenados con los datos transformadas del apartado anterior.

El apartado *c*) se calcula como los ejercicios 1 y 2. Es recomendable utilizar los valores primero y último de la lista del apartado *a*).

El apartado *d*) incluye diferentes cálculos que indicamos a continuación. Recuerda que para hacer previsiones hacia el futuro hacemos la hipótesis de que las magnitudes evolucionan al ritmo del incremento medio anual que tomamos para hacer los cálculos.

Hay que hacer las previsiones de los impuestos en términos reales, multiplicando la última fecha conocida en términos reales y multiplicándola por el incremento medio anual tantas veces como años pasen.

Hay que hacer también las previsiones del IPC de los años venideros, aplicando a la última fecha publicada el incremento medio anual obtenido de los IPC de los años considerados.

También necesitamos convertir las previsiones de los resultados anteriores en términos reales, en términos relativos o monedas corrientes utilizando los valores del IPC del año base y las estimaciones del IPC de los años venideros.

Ejercicio 8

En este ejercicio vamos a calcular la pérdida o ganancia de poder adquisitivo de un salario del que conocemos los valores.

También hay que obtener los valores del IPC de los mismos años de la web del INE.

Con las dos listas hay que calcular los incrementos anuales mediante los índices encadenados.

Recuerda que para calcular las variaciones del poder adquisitivo es necesario hacer esta operación:

$$\text{Ganancia o pérdida de poder adquisitivo} = \frac{(1 + \Delta_{salari})}{(1 + \Delta_{IPC})}$$

Luego se completa la tabla con la columna equivalente, calculada con los datos inicial y final del período.

Acabaremos con esta tabla completada, donde ya hemos indicado los incrementos anuales antes mencionados:

	2008	**2009**	**2010**	**2011**	**TOTAL**
Incremento salarial	+1,4 %	+7,6 %	+1,6 %	-3,2 %	
Incremento IPC	+4,1 %	-0,3 %	+1,8 %	+2,1 %	
Pérdida o ganancia poder adquisitivo					

Ejercicio 9

Para hacer este ejercicio solo tienes que hacer el mismo esquema del ejercicio 8.

Ejercicio 10

Para empezar hay que hacer un cambio de base y encontrar todos los valores del IPC en la misma base. Llenaremos la tabla con los nuevos valores encontrados, tal como se hizo en el ejercicio 4.

	2000	2001	2002	2003	2004
IPC base 1992	131	135	139		
IPC base 2002			103	106	109

a) Calcula el incremento anual, medio y total del precio del ordenador en términos reales.

Dado que el análisis del precio del ordenador nos lo piden en términos reales, primero es necesario hacer la conversión del precio de la computadora en moneda constante del año 2000 (deflactación de la serie de precios de la computadora). Podemos ayudarnos de las siguientes tablas para dar los resultados de cada año con más claridad.

Año	Precio ordenador (términos nominales)	Precios ordenador (términos reales 2000)
2000	1300	1300
2001	1275	$1275 \cdot \frac{IPC_{1992}^{2000}}{IPC_{1992}^{2001}} =$ ....
2002	1250	$1250 \cdot \frac{IPC_{1992}^{2000}}{IPC_{1992}^{2002}} =$
2003	1100	
2004	950	

Para calcular el incremento anual del precio del ordenador en términos reales, operamos en la tabla siguiente mediante índices encadenados.

Año	Precio ordenador (términos nominales)	Precio ordenador (términos reales 2000)	Índice	Interpretación: Incremento anual
2000	1300	1300	-----	
2001	1275	1237,2	$I_{2000}^{2001} = \frac{1237,2}{1300} = \ldots\ldots$	Ha disminuido un ....%
2002	1250	1178,06		
2003	1100	1007,34		
2004	950	846,02		

Y finalmente, con los datos de la tabla, calcularemos el incremento total y medio del período.

b) Si seguimos esta evolución, estimamos el precio que podría tener el ordenador en 2008.

Para hacer este apartado, consideremos que nos piden la estimación en términos corrientes del precio, suponiendo que no varía el comportamiento del IPC ni la evolución del precio del ordenador en términos reales, que hemos analizado en el apartado anterior.

Para hacer estas estimaciones necesitamos el incremento medio de las dos series y con estos datos, podremos estimar el valor del IPC el año 2008.

Haremos las mismas operaciones con la serie de los precios de los ordenadores en términos reales y, por último, hay que pasar el resultado a términos corrientes en moneda del año 2008.

Ayudas Tipo 2

Ejercicio 1

En el apartado *a*), para calcular los índices, proponemos llenar esta tabla:

	Número total alumnos matriculados	Índice Base 2005
Curso 2005/2006	12676	$I_{2005}^{2005} = 1$
Curso 2006/2007	12928	$I_{2005}^{2006} = \frac{12928}{12676} = 1{,}01988009$
Curso 2007/2008	13159	$I_{2005}^{2007} = \frac{13159}{12676} =$
Curso 2008/2009	13210	$I_{2005}^{2008} = \frac{13210}{12676} =$
Curso 2009/2010	13904	$I_{2005}^{2009} =$
Curso 2010/2011	14702	$I_{2005}^{2010} =$

Ahora hay que interpretar los resultados de la columna de la derecha.

En el apartado *b*) nos piden los índices encadenados y también podemos llenar la siguiente tabla:

	Número total alumnos matriculados	Índices encadenados
Curso 2005/2006	12676	---
Curso 2006/2007	12928	$I_{2005}^{2006} = \frac{12928}{12676} = 1{,}01988009$
Curso 2007/2008	13159	$I_{2006}^{2007} = \frac{13159}{12928} =$
Curso 2008/2009	13210	$I_{2007}^{2008} =$
Curso 2009/2010	13904	$I_{2008}^{2009} =$
Curso 2010/2011	14702	$I_{2009}^{2010} =$

Interpretaremos también los resultados de la columna de la derecha. Como los índices son todos mayores que 1, indica que la serie siempre aumenta pero a diferente ritmo, según el año que analicemos.

En el apartado *c*) nos piden el incremento total del período $I_{2005}^{2010} = \dfrac{14702}{12676} = 1{,}16$.

Y el incremento medio anual con la raíz quinta de este cociente.

En el apartado *d*), para hacer previsiones, partimos de la hipótesis de que el incremento medio anual que hemos obtenido en el apartado anterior será una estimación del incremento anual de los años que están por venir, y a partir del último dato conocido calcularemos las cantidades de alumnos que podremos esperar.

Así, para estimar la cantidad de alumnos que podemos esperar que se matricule en el curso 2011/2012, será de 14.702 · 1,03 = 15.143 alumnos.

Con el mismo razonamiento podemos hacer el resto de estimaciones.

Ejercicio 2

Este ejercicio es igual que el ejercicio 1, pero obtendremos algunos índices por encima de1 y otros por debajo. Dejemos la interpretación para los lectores.

En el apartado *a*) hay que hacer los cálculos que sugerimos en la columna de la derecha de la siguiente tabla:

	Número total visitantes	Índice Base 2000
2000	10252799	$I_{2000}^{2000} = 1$
2001	10002517	$I_{2000}^{2001} = \dfrac{10002517}{10252799} =$
2002	9661493	$I_{2000}^{2002} = \dfrac{9661493}{10252799} = 0{,}94232736$
2003	10296382	$I_{2000}^{2003} = \dfrac{10296382}{10252799} = 1{,}00425084$
2004	11134880	$I_{2000}^{2004} = \dfrac{11134880}{10252799} =$
2005	10743480	$I_{2000}^{2005} = \dfrac{10743480}{10252799} = 1{,}04785825$

2006	10979470	$I_{2000}^{2006} = \dfrac{10979470}{10252799} =$
2007	10864738	$I_{2000}^{2007} =$
2008	10222818	$I_{2000}^{2008} =$
2009	9952606	$I_{2000}^{2009} =$

Hay que hacer la interpretación de estos resultados, especialmente prestar atención a la de los índices menores que 1.

Para calcular los índices encadenados del apartado *b*) también sugerimos los cálculos por realizar en la última columna de la siguiente tabla:

	Número total visitantes	Índices encadenados
2000	10252799	
2001	10002517	$I_{2000}^{2001} = \dfrac{10002517}{10252799} = 0{,}97558891$
2002	9661493	$I_{2001}^{2002} = \dfrac{9661493}{10002517} =$
2003	10296382	$I_{2002}^{2003} = \dfrac{10296382}{9661493} =$
2004	11134880	$I_{2003}^{2004} = \dfrac{}{10296382} = 1{,}08143618$
2005	10743480	$I_{2004}^{2005} = \dfrac{10743480}{} = 0{,}96484919$
2006	10979470	$I_{2005}^{2006} = ----= 1{,}02196588$
2007	10864738	$I_{2006}^{2007} =$
2008	10222818	$I_{2007}^{2008} =$
2009	9952606	$I_{2008}^{2009} =$

Dejemos las interpretaciones de los resultados al lector.

En el apartado *c*) nos piden el incremento total. Hay que obtener el siguiente índice I_{2000}^{2009} y con el resultado hacer la raíz que nos permita averiguar el valor del incremento medio anual del –0,33 %.

En el apartado *d*) nos piden previsiones que calcularemos a partir del últim dato conocido (9.952.606) y aplicando reiteradamente el factor correspondiente al incremento medio anual del –0,33 % (0,9967).

Ejercicio 3

Este ejercicio trabaja los mismos conceptos que los anteriores, pero empezando por incrementos porcentuales que permiten convertirlos en índice encadenados directamente. En la tabla siguiente se indican algunos valores y os invitamos a completarla:

Año	Variaciones del volumen de ventas (%)	Índices encadenados
2006	-3,13	$I_{2005}^{2006} = 0{,}9687$
2007	-2,15	$I_{2006}^{2007} =$
2008	+2,12	$I_{2007}^{2008} = 1{,}0212$
2009	+3,15	$I_{2008}^{2009} =$
2010	+4,12	$I_{2009}^{2010} = 1{,}0412$
2011	+4,31	$I_{2010}^{2011} =$

Para calcular los índices de base 2005, multiplicaremos los índices anteriores (aquí tienes calculados algunos de ellos. Calcula tú el resto).

$$I_{2005}^{2007} = I_{2005}^{2006} \cdot I_{2006}^{2007} = 0{,}9687 \cdot 0{,}9785 = 0{,}9479$$

$$I_{2005}^{2008} = I_{2005}^{2006} \cdot I_{2006}^{2007} \cdot I_{2007}^{2008} = 0{,}9687 \cdot 0{,}9785 \cdot 1{,}0212 = 0{,}9680$$

$$I_{2005}^{2009} = I_{2005}^{2006} \cdot I_{2006}^{2007} \cdot I_{2007}^{2008} \cdot I_{2008}^{2009} =$$

$$I_{2005}^{2010} = I_{2005}^{2006} \cdot I_{2006}^{2007} \cdot I_{2007}^{2008} \cdot I_{2008}^{2009} \cdot I_{2009}^{2010} =$$

$$I_{2005}^{2011} = I_{2005}^{2006} \cdot I_{2006}^{2007} \cdot I_{2007}^{2008} \cdot I_{2008}^{2009} \cdot I_{2009}^{2010} \cdot I_{2010}^{2011} =$$

$$= 0{,}9687 \cdot 0{,}9785 \cdot 1{,}0212 \cdot 1{,}0315 \cdot 1{,}0412 \cdot 1{,}0431 = 1{,}0844$$

En el apartado *b*) nos piden el incremento total a partir de los factores de los índices, que es el último cálculo que hemos visto y así, el incremento medio anual es:

$$\sqrt[6]{0{,}9687 \cdot 0{,}9785 \cdot 1{,}0212 \cdot 1{,}0315 \cdot 1{,}0412 \cdot 1{,}0431} = \sqrt[6]{1{,}0844} = 1{,}0136$$

que podemos interpretar como que el aumento total es equivalente a un incremento anual constante del 1,36 %.

En el aparato *c*) nos piden previsiones y hay que reflexionar si este supuesto es posible.

Ejercicio 4

En este ejercicio nos presentan el proceso de cálculo de un cambio de base de los índices de precios (IPC).

A tal fin hay que considerar los datos que presentamos en esta tabla y plantear proporcionalidades como las que mostramos a continuación para llenar las casillas sombreadas en gris.

	2002	2003	2004	2005	2006	2007	2008	2009	2010
IPC Base 2001	117,624	117,624	117,624	117,624	117,624				
IPC Base 2006	88,024	90,699			100	102,787	106,976	106,668	108,588

$$IPC_{2001}^{2002} = 103{,}538 \rightarrow IPC_{2006}^{2002} = \boldsymbol{x}$$

$$IPC_{2001}^{2006} = 117{,}624 \rightarrow IPC_{2006}^{2006} = 100$$

$$IPC_{2006}^{2002} = \frac{IPC_{2001}^{2002} \cdot IPC_{2006}^{2006}}{IPC_{2001}^{2006}} = \frac{103{,}538 \cdot 100}{117{,}624} = 103{,}538 \frac{100}{117{,}624} =$$

$= 103{,}538 \cdot 0{,}8502 = 88{,}025$

$$IPC_{2006}^{2003} = \frac{IPC_{2001}^{2003} \cdot IPC_{2006}^{2006}}{IPC_{2001}^{2006}} = \frac{106{,}684 \cdot 100}{117{,}624} = 106{,}684 \frac{100}{117{,}624} =$$

$= 106{,}684 \cdot 0{,}8502 = 90{,}699$

Y así, hasta completar la tabla.

Es evidente, que para obtener los valores del IPC del año «y» en base 2006 hemos multiplicado el IPC del año «y» en base 2001 por la fracción $\frac{IPC_{2006}^{2006}}{IPC_{2001}^{2006}}$ el valor de la que hemos reseñado en rojo y es lo que denominaremos «enlace técnico».

Aunque no se contempla en el ejercicio, si dividimos los valores del IPC de la segunda tabla por este «enlace técnico», también podríamos obtener los IPC del período 2006-2010 en base 2001.

Ejercicio 5

Como en las ayudas de tipo 1 ya hemos presentado las fórmulas por emplear en cada uno de los casos y el significado de la notación, pondremos un ejemplo hecho de cada una de ellas:

Índice de precios de Laspeyres

$$L_{precios}{}_{2008}^{2009} = \frac{\sum_i p_{i2009} \cdot q_{i2008}}{\sum_i p_{i2008} \cdot q_{i2008}} = \frac{14 \cdot 100 + 8 \cdot 50 + 10 \cdot 20}{12 \cdot 100 + 10 \cdot 50 + 5 \cdot 20} = \frac{2000}{1800} = 1{,}111$$

Índice de cantidades de Laspeyres

$$L_{cantidades}{}_{2008}^{2009} = \frac{\sum_i q_{i2009} \cdot p_{i2008}}{\sum_i p_{i2008} \cdot q_{i2008}} = \frac{112 \cdot 12 + 65 \cdot 10 + 10 \cdot 5}{12 \cdot 100 + 10 \cdot 50 + 5 \cdot 20} = \frac{2044}{1800} = 1{,}136$$

Índice de precios de Paasche

$$P_{precios}{}_{2008}^{2009} = \frac{\sum_i p_{i2009} \cdot q_{i2009}}{\sum_i p_{i2008} \cdot q_{i2009}} = \frac{14 \cdot 112 + 8 \cdot 65 + 10 \cdot 10}{12 \cdot 112 + 10 \cdot 65 + 5 \cdot 10} = \frac{2188}{2044} = 1{,}070$$

Índice de cantidades de Paasche

$$P_{cantidades}{}_{2008}^{2009} = \frac{\sum_i p_{i2009} \cdot q_{i2009}}{\sum_i q_{i2008} \cdot p_{i2009}} = \frac{14 \cdot 112 + 8 \cdot 65 + 10 \cdot 10}{14 \cdot 100 + 8 \cdot 50 + 10 \cdot 20} = \frac{2188}{2000} = 1{,}094$$

Ejercicio 6

La primera operación permitirá hacer el cambio a euros de la cantidad de compra: 96.913,20 euros en moneda corriente del año 1998, y para transformarla en moneda del año 2006, haremos la siguiente operación:

$$96913,20 \cdot \frac{IPC_{2001}^{2006}}{IPC_{2001}^{1998}} = 96913,20 \cdot \frac{117,624}{91,223} = 124961,01€$$

Los valores del IPC se encuentran en la web del INE con el consecuente cambio de base, como aprendimos en el ejercicio 4.

Como en el enunciado se dice que la hemos vendido por 240.000 euros, vamos a calcular los beneficios en términos relativos, a partir del concepto de índice: $\frac{240000}{124961,01}$. Interpreta el resultado.

Ejercicio 7

Para deflactar la serie (pasarla a términos reales del año 2006) hay que hacer las operaciones que se reflejan en la siguiente tabla. A tal fin, es necesario que consultemos los valores del IPC de estos años. Tomaremos las medias anuales del IPC general.

Año	Importe impuesto municipal (términos nominales)	Importe impuesto municipal (términos reales 2006)
2006	503,24	503,24
2007	515,65	$515,65 \cdot \frac{IPC_{2006}^{2006}}{IPC_{2006}^{2007}} = 515,65 \cdot \frac{100}{102,787} = 501,67$
2008	536,73	$536,73 \cdot \frac{IPC_{2006}^{2006}}{IPC_{2006}^{2008}} = 536,73 \cdot \frac{100}{106,976} =$
2009	578,84	$578,84 \cdot \frac{IPC_{2006}^{2006}}{IPC_{2006}^{2009}} = 578,84 \cdot \frac{100}{106,668} =$
2010	584,42	$584,42 \cdot \frac{IPC_{2006}^{2006}}{IPC_{2006}^{2010}} = 584,42 \cdot \frac{100}{108,588} =$

En el apartado *b*) nos piden los índices encadenados con los valores en términos reales que presentamos en la siguiente tabla.

Año	Importe impuesto municipal (términos nominales)	Importe impuesto municipal (términos reales 2006)	Índice	
2006	503,24	503,24	-----	
2007	515,65	501,67	$I_{2006}^{2007}=\dfrac{501,67}{503,24}=0,997$	Ha disminuido un 0,3 %
2008	536,73	501,73	$I_{2007}^{2008}=\dfrac{501,73}{501,67}$	Podemos considerar que es constante
2009	578,84	542,66	$I_{2008}^{2009}=$	
2010	584,42	538,20	$I_{2009}^{2010}=$	Ha disminuido un ...%

En el apartado *c*) nos piden el incremento total y el incremento medio anual pero disponemos de las magnitudes de la columna correspondiente (tercera de la tabla anterior).

Para calcular el incremento total del período, interpretaremos el índice:

$$I_{2006}^{2010}=\frac{538,20}{503,24}$$

y para calcular el incremento medio anual, calcularemos la raíz cuarta del resultado anterior.

Para abordar el apartado *d*), y como se trata de hacer estimaciones, supondremos que los fenómenos evolucionarán con el ritmo que podamos interpretar del incremento medio anual de cada una.

Calcularemos, en primer lugar, lo que corresponde a los IPC de los años 2006 al 2010 con los datos de la web del INE. Hemos obtenido un incremento del 2,08 % anual y con este dato y el IPC del año 2010 calcularemos los IPC de los años 2011, 2012 y 2013.

También estimaremos el importe del impuesto en términos reales de estos años, aplicando el incremento medio anual (1,7 %) sobre el valor de este importe del año 2010, es decir, 538,20 euros en términos reales del 2006.

Como estos resultados están expresados en términos reales del año 2006, hay que convertirlos a términos nominales, utilizando los IPC que acabamos de estimar también.

Indicamos las operaciones del primer resultado:

$$\text{Año 2011} \rightarrow 547,25\cdot\frac{IPC_{2006}^{2011}}{IPC_{2006}^{2006}}=547,25\cdot\frac{110,847}{100}=606,61$$

Y el resto lo calculamos de la misma manera.

Ejercicio 8

En este ejercicio se trata de estudiar la pérdida o ganancia de poder adquisitivo de un salario.

Habrá que calcular los incrementos anuales de los salarios como se indica en la siguiente tabla:

Año	Nómina mensual (€)	Índice	
2007	2034,75	------	-------
2008	2062,13	$I_{2007}^{2008} = \frac{2062,13}{2034,75} = 1,014$	Ha aumentado 1,4 %
2009	2218,61	$I_{2008}^{2009} = \frac{\quad}{2062,13} = 1,076$	Ha aumentado ...%
2010	2253,67	$I_{2009}^{2010} = \frac{2253,67}{\quad} =$	Ha aumentado 1,6 %
2011	2181,75	$I_{2010}^{2011} = ---- = 0,968$	Ha disminuido ...%

También calcularemos los incrementos anuales del IPC de los años correspondientes:

Año	IPC Base 2006	Índice	
2007	$IPC_{2006}^{2007} = 102,787$	------	-------
2008	$IPC_{2006}^{2008} = 106,976$	$I_{2007}^{2008} = \frac{106,976}{102,787} = 1,041$	Ha aumentado 4,1 %
2009	$IPC_{2006}^{2009} = 106,668$	$I_{2008}^{2009} = \frac{106,668}{106,976} =$	Ha aumentado ...%
2010	$IPC_{2006}^{2010} = 108,588$	$I_{2009}^{2010} = \frac{108,588}{\quad} =$	Ha aumentado 1,8 %
2011	$IPC_{2006}^{2011} = 110,847$	$I_{2010}^{2011} = \frac{\quad}{108,588} =$	Ha disminuido ...%

Con estos resultados hay que rellenar las casillas en gris de la siguiente tabla:

	2008	2009	2010	2011
Incremento salarial	1,014	1,076	1,016	0,968
Incremento IPC	1,041	0,997	1,018	1,021
Pérdida o ganancia poder adquisitivo	0,974			

Indicamos cómo obtener el resultado que hemos puesto en la tabla.

Año 2008 → $\frac{1{,}014}{1{,}041} = 0{,}974$, de la misma manera podemos llenar el resto de casillas.

Con los valores de los IPC y los salarios podemos obtener el incremento total del período de cada una de las magnitudes. Asimismo, añadir los resultados en la columna de la derecha de la tabla:

Indicamos todos los resultados obtenidos para comprobar:

	2008	2009	2010	2011	TOTAL
Incremento salarial	+1,4 %	+7,6 %	+1,6 %	–3,2 %	+7,2 %
Incremento IPC	+4,1 %	–0,3 %	+1,8 %	+2,1 %	+7,8 %
Pérdida o ganancia poder adquisitivo	–2,6 %	+7,9 %	–0,2 %	–5,2 %	–0,6 %

Ejercicio 9

Para hacer este ejercicio tienes que seguir las pautas del ejercicio 8.

Ejercicio 10

Para empezar, hay que hacer un cambio de base y encontrar todos los valores del IPC en la misma base. Llenaremos la tabla con los nuevos valores encontrados, como se hizo en el ejercicio 4.

	2000	2001	2002	2003	2004
IPC base 1992	131	135	139		
IPC base 2002			103	106	109

Pueden plantearse por proporcionalidad, como aquí se puede ver para el año 2003:

$$IPC_{2002}^{2002} = 103 \rightarrow IPC_{2002}^{2003} = 106$$

$$IPC_{1992}^{2002} = 139 \rightarrow IPC_{1992}^{2003} = X$$ y calcular el dato desconocido.

a) Calcula el incremento anual, medio y total del precio del ordenador en términos reales.

Dado que el análisis del precio del ordenador nos lo piden en términos reales, es necesario primero hacer la conversión de su precio en moneda constante del año 2000 (deflactación de la serie de precios del ordenador) como hemos empezado en la tabla siguiente;

Año	Precio ordenador (términos nominales)	Precio ordenador (términos reales 2000)
2000	1300	1300
2001	1275	$1275\cdot\frac{IPC_{1992}^{2000}}{IPC_{1992}^{2001}}=1275\cdot\frac{131}{135}=1237{,}2$
2002	1250	$1250\cdot\frac{IPC_{1992}^{2000}}{IPC_{1992}^{2002}}=$
2003	1100	$1100\cdot\frac{IPC_{1992}^{2000}}{IPC_{1992}^{2003}}=$
2004	950	$950\cdot\frac{IPC_{1992}^{2000}}{IPC_{1992}^{2004}}=$

Para calcular el incremento anual del precio del ordenador en términos reales, operamos en la tabla siguiente mediante índices encadenados (intenta acabarla).

Año	Precio ordenador (términos nominales)	Precio ordenador (términos reales 2000)	Índice	Interpretación: Incremento anual
2000	1300	1300	-----	
2001	1275	1237,2	$I_{2000}^{2001}=\frac{1237{,}2}{1300}=0{,}9517$	Ha disminuido un 4,83 %
2002	1250	1178,06		Ha disminuido un 4,78 %
2003	1100	1007,34	$I_{2002}^{2003}=\frac{1007{,}34}{1178{,}06}=0{,}8551$	
2004	950	846,02		

Para conocer el incremento total y medio del período, calcularemos I_{2000}^{2004} y la raíz correspondiente de índice 4.

a) Si seguimos esta evolución, valoremos el precio que podría tener el ordenador en 2008.

Para hacer este apartado, consideremos que nos piden la estimación en términos corrientes del precio, suponiendo que no varía el comportamiento del IPC ni la evolución del precio del ordenador en términos reales que hemos analizado en el apartado anterior.

Para hacer estas estimaciones necesitamos el incremento medio de las dos series y, con estos datos, podemos calcular el valor del IPC en el año 2008:

$$IPC_{1992}^{2008} = IPC_{1992}^{2004} \cdot 1{,}0294^4 = 147{,}10 \cdot 1{,}0294^4 = 165{,}18$$

Haremos las mismas operaciones con la serie de los precios de los ordenadores en términos reales.

Precio ordenador para el año 2008 en términos reales del 2000 = $846{,}02 \cdot 0{,}8982^4$ = 550,65 euros.

Y finalmente, hay que pasar a términos corrientes en moneda del 2008.

Soluciones

Ejercicio 1

A continuación presentamos el volumen total de alumnos matriculados en la Universitat Jaume I en los últimos años.

	Número total de alumnos matriculados
Curso 2005/2006	12676
Curso 2006/2007	12928
Curso 2007/2008	13159
Curso 2008/2009	13210
Curso 2009/2010	13904
Curso 2010/2011	14702

a) Calcula los índices para cada año, tomando como año de referencia el 2005 (hará referencia al curso 2005/2006). Interpreta el resultado.
b) Calcula los índices encadenados de esta serie. Interpreta los resultados.
c) Calcula el incremento total e incremento medio anual de este período, a partir de las cantidades originales y a partir de los índices encadenados.
d) Haz previsiones para la matrícula de los cursos 2011/2012 y 2012/2013, si consideramos que no habrá cambios significativos en su comportamiento.

Solución

a) Calcula los índices para cada año, tomando como año de referencia el 2005 (hará referencia al curso 2005/2006). Interpreta el resultado.

Tomaremos como fecha de referencia, los 12.676 alumnos matriculados el curso 2005/2006 y en la siguiente tabla indicaremos los cocientes correspondientes a los índices que queremos calcular:

	Número total alumnos matriculados	Índice Base 2005
Curso 2005/2006	12676	$I_{2005}^{2005} = 1$
Curso 2006/2007	12928	$I_{2005}^{2006} = \dfrac{12928}{12676} = 1{,}01988009$
Curso 2007/2008	13159	$I_{2005}^{2007} = \dfrac{13159}{12676} = 1{,}0381035$
Curso 2008/2009	13210	$I_{2005}^{2008} = \dfrac{13210}{12676} = 1{,}04212685$
Curso 2009/2010	13904	$I_{2005}^{2009} = \dfrac{13904}{12676} = 1{,}09687599$
Curso 2010/2011	14702	$I_{2005}^{2010} = \dfrac{14702}{12676} = 1{,}1598296$

Para interpretar estos datos consideraremos que estos cocientes comparan dos magnitudes en términos relativos y son cantidades que no vienen expresadas en ninguna unidad, sino que podemos interpretarlas como porcentajes. Así, si redondeamos los resultados con dos cifras decimales podremos decir:

$I_{2005}^{2006} = 1{,}02$ nos indica que el número de alumnos del curso 2006 es un 2 % superior al del año 2005. También podríamos expresarlos así $I_{2005}^{2006} = \dfrac{12928}{12676} \cdot 100 = 102$ %.

$I_{2005}^{2007} = 1{,}04$ nos indica que el número de alumnos del curso 2007 es un 4 % superior al del 2005.

$I_{2005}^{2008} = 1{,}04$ nos indica que el número de alumnos del curso 2008 es un 4 % superior al del año 2005. Indica un cierto estacionamiento en el aumento del número de alumnos.

$I_{2005}^{2009} = 1{,}10$ nos indica que el número de alumnos del curso 2009 es un 10% superior al del 2005.

$I_{2005}^{2010} = 1{,}16$ nos indica que el número de alumnos del curso 2010 es un 16 % superior al del 2005.

Estos índices dan idea del aumento del número total de alumnos de la UJI, considerando siempre como referencia el número de alumnos del curso 2005 que podremos considerar que sería del 100 %.

b) Calcula los índices encadenados de esta serie. Interpreta los resultados.

Para calcular estos índices utilizaremos la siguiente tabla, que podremos comparar con la tabla del apartado anterior:

	Número total alumnos matriculados	Índice encadenado
Curso 2005/2006	12676	
Curso 2006/2007	12928	$I_{2005}^{2006} = \frac{12928}{12676} = 1{,}01988009$
Curso 2007/2008	13159	$I_{2006}^{2007} = \frac{13159}{12928} = 1{,}01786819$
Curso 2008/2009	13210	$I_{2007}^{2008} = \frac{13210}{13159} = 1{,}00387567$
Curso 2009/2010	13904	$I_{2008}^{2009} = \frac{13904}{13210} = 1{,}05253596$
Curso 2010/2011	14702	$I_{2009}^{2010} = \frac{14702}{13904} = 1{,}05739356$

Como se puede ver en estos índices, la cantidad que tomamos como referencia es el número de alumnos matriculados en el año anterior, por lo que podemos denominarlos «índices encadenados» y nos permitirá ver el crecimiento año tras año.

$I_{2005}^{2006} = 1{,}02$ nos indica que el número de alumnos del curso 2006 es un 2 % superior al del 2005.

$I_{2006}^{2007} = 1{,}02$ nos indica que el número de alumnos del curso 2007 es un 2 % superior al del año 2006.

$I_{2007}^{2008} = 1{,}004$ nos indica que el número de alumnos del curso 2008 es un 0,4 % superior al del año 2007.

$I_{2008}^{2009} = 1{,}05$ nos indica que el número de alumnos del curso 2009 es un 5 % superior al del año 2008.

$I_{2009}^{2010} = 1{,}06$ nos indica que el número de alumnos del curso 2010 es un 6 % superior al del año 2009.

Estos índices dan idea del aumento del número total de alumnos de la UJI, pero detallando la evolución por años.

c) Calcula el incremento total e incremento medio anual de este período, a partir de las cantidades originales y a partir de los índices encadenados.

Para calcular el incremento total basándonos en las cantidades originales, tan solo hay que considerar el número de alumnos de los primeros y últimos cursos, para compararlos. Este concepto corresponde al último índice calculado en el apartado *a*). Así:

$I_{2005}^{2010} = \frac{14702}{12676} = 1,16$ que ya hemos comentado que nos indica que el número de alumnos se ha incrementado en un 16 % en el período estudiado.

Para obtener el incremento medio anual, habrá que considerar la raíz con un índice que viene dado por el número de incrementos que contiene el período estudiado. Notamos que corresponde al número de años menos uno. Así, en nuestro ejercicio:

$$\sqrt[5]{\frac{14702}{12676}} = \sqrt[5]{1,1598296} = 1,03$$

y podremos interpretar que este crecimiento total del 16 % sería equivalente a un crecimiento constante anual del 3 %.

Reseñamos que en estos cálculos ya hechos hemos utilizado el número de alumnos del primer y último curso que consideramos en el período.

En otras magnitudes podemos no conocer estas cantidades, pero sí los incrementos anuales o índices encadenados que nosotros hemos obtenido en el apartado *b*). En este caso también podremos calcular este incremento total e incremento medio anual. Veamos:

$$I_{2005}^{2010} = I_{2005}^{2006} \cdot I_{2006}^{2007} \cdot I_{2007}^{2008} \cdot I_{2008}^{2009} \cdot I_{2009}^{2010} =$$

$$= \frac{12928}{12676} \cdot \frac{13159}{12928} \cdot \frac{13210}{13159} \cdot \frac{13904}{13210} \cdot \frac{14702}{13904} = \frac{14702}{12676} =$$

$= 1,01988009 \cdot 1,01786819 \cdot 1,00387567 \cdot 1,05253596 \cdot 1,05739356 = 1,1598296$ que nos permite interpretar que en el período considerado el número de alumnos ha aumentado un 16 %.

Para calcular el incremento medio, calcularemos pues la raíz quinta del producto de los índices (el índice de la raíz coincide con el número de índices que forman el producto).

$$\sqrt[5]{I_{2005}^{2006} \cdot I_{2006}^{2007} \cdot I_{2007}^{2008} \cdot I_{2008}^{2009} \cdot I_{2009}^{2010}} =$$

$$\sqrt[5]{1,01988009 \cdot 1,01786819 \cdot 1,00387567 \cdot 1,05253596 \cdot 1,05739356} = \sqrt[5]{1,1598296} = 1,03$$

que interpretaremos como un incremento anual constante del 3 %.

d) Haz previsiones para la matrícula de los cursos 2011/2012 y 2012/2013, si consideramos que no habrá cambios significativos en su comportamiento.

Para hacer previsiones hay que partir de la hipótesis de que el incremento medio anual que hemos obtenido en el apartado anterior, podría ser una estimación del incremento anual de los años que están por venir y que, a partir del último dato conocid, calcularemos las cantidades de alumnos que podremos esperar.

Así, para estimar la cantidad de alumnos que podemos esperar que se matricule en el curso 2011/2012, será de $14.702 \cdot 1{,}03 = 15.143$ alumnos.

Y para estimar la cantidad de alumnos que podremos esperar para el curso 2012/2013 hay que considerar que pasarán dos años desde la fecha de la última cantidad de alumnos real $14.702 \cdot 1{,}03^2 = 15.507$ alumnos.

Notamos que por previsiones o estimaciones partiremos del último dato real y lo multiplicaremos por el incremento medio anual que hemos calculado previamente, y elevándolo al número de años o períodos que están por venir.

Ejercicio 2

En la siguiente tabla se muestran los datos del INI que hacen referencia al total de visitantes a los parques nacionales de España, en los años indicados.

Naturaleza y biodiversidad	
Zonas protegidas	
Número de visitantes por nacionalidades y período	
Unidades: número de personas	
	Total
2000	10252799
2001	10002517
2002	9661493
2003	10296382
2004	11134880
2005	10743480
2006	10979470
2007	10864738
2008	10222818
2009	9952606
Fuente: Ministerio de Medio Ambiente y Medio Rural y Marino. Red de Parques Naturales Copyright INE 2011	

a) Calcula los índices para cada año, tomando como año de referencia el 2000 e interpreta los resultados.
b) Calcula los índices encadenados de esta serie. Interpreta los resultados.
c) Calcula el incremento total e incremento medio anual de este período, a partir de las cantidades originales y a partir de los índices encadenados.
d) Haz previsiones del número de visitantes de los parques considerados para los años 2010, 2011 y 2012, si consideramos que no hubiera cambios significativos en el comportamiento de la afluencia.

Fuente: INE

Solución

a) Calcula los índices para cada año, tomando como año de referencia el 2000 e interpreta los resultados.

Tomaremos como fecha de referencia, los 10.252.799 visitantes que recorrieron los parques naturales de España en su totalidad en el año 2000 y en la siguiente tabla indicaremos los cocientes correspondientes a los índices que queremos calcular:

	Número total visitantes	Índice Base 2000
2000	10252799	$I_{2000}^{2000} = 1$
2001	10002517	$I_{2000}^{2001} = \frac{10002517}{10252799} = 0{,}97558891$
2002	9661493	$I_{2000}^{2002} = \frac{9661493}{10252799} = 0{,}94232736$
2003	10296382	$I_{2000}^{2003} = \frac{10296382}{10252799} = 1{,}00425084$
2004	11134880	$I_{2000}^{2004} = \frac{11134880}{10252799} = 1{,}08603319$
2005	10743480	$I_{2000}^{2005} = \frac{10743480}{10252799} = 1{,}04785825$
2006	10979470	$I_{2000}^{2006} = \frac{10979470}{10252799} = 1{,}07087538$
2007	10864738	$I_{2000}^{2007} = \frac{10864738}{10252799} = 1{,}05968507$
2008	10222818	$I_{2000}^{2008} = \frac{10222818}{10252799} = 0{,}99707582$
2009	9952606	$I_{2000}^{2009} = \frac{9952606}{10252799} = 0{,}97072087$

Para interpretar estos datos consideraremos que estos cocientes comparan dos magnitudes en términos relativos y podemos interpretarlas como porcentajes. A diferencia del apartado anterior, esta magnitud disminuye y crece según de qué período sean los datos. Así, si redondeamos los resultados con dos cifras decimales podremos decir:

$I_{2000}^{2001} = 0{,}98$ nos indica que el número de visitantes ha disminuido un 2 % del año 2000 al 2001.

$I_{2000}^{2002} = 0{,}94$ nos indica que el número de visitantes ha disminuido un 6 % del año 2000 al 2002.

$I_{2000}^{2003} = 1{,}004$ nos indica que el número de visitantes ha aumentado un 0,4 % del año 2000 al 2003.

$I_{2000}^{2004} = 1{,}09$ nos indica que el número de visitantes ha aumentado un 9 % del año 2000 al 2004.

$I_{2000}^{2005} = 1{,}05$ nos indica que el número de visitantes ha aumentado un 5 % del año 2000 al 2005.

$I_{2000}^{2006} = 1{,}07$ nos indica que el número de visitantes ha aumentado un 7 % del año 2000 al 2006.

$I_{2000}^{2007} = 1{,}06$ nos indica que el número de visitantes ha aumentado un 6 % del año 2000 al 2007.

$I_{2000}^{2008} = 0{,}997$ nos indica que el número de visitantes ha aumentado un 0,3 % del año 2000 al 2008.

$I_{2000}^{2009} = 0{,}97$ nos indica que el número de visitantes ha aumentado un 5 % del año 2000 al 2009.

Estos índices dan idea de las variaciones en el número de visitantes a los parques de España, a pesar de que tan solo consideran los valores del comienzo y la finalización del período referido y no se reflejan las fluctuaciones dentro del período.

b) Calcula los índices encadenados de esta serie. Interpreta los resultados.

Para calcular estos índices, utilizaremos la siguiente tabla, que podremos compararla con la del apartado anterior:

	Número total visitantes	Índice encadenados
2000	10252799	
2001	10002517	$I_{2000}^{2001} = \frac{10002517}{10252799} = 0{,}97558891$
2002	9661493	$I_{2001}^{2002} = \frac{9661493}{10002517} = 0{,}96590618$
2003	10296382	$I_{2002}^{2003} = \frac{10296382}{9661493} = 1{,}06571334$
2004	11134880	$I_{2003}^{2004} = \frac{11134880}{10296382} = 1{,}08143618$
2005	10743480	$I_{2004}^{2005} = \frac{10743480}{11134880} = 0{,}96484919$
2006	10979470	$I_{2005}^{2006} = \frac{10979470}{10743480} = 1{,}02196588$
2007	10864738	$I_{2006}^{2007} = \frac{10864738}{10979470} = 0{,}98955032$
2008	10222818	$I_{2007}^{2008} = \frac{10222818}{10864738} = 0{,}94091712$
2009	9952606	$I_{2008}^{2009} = \frac{9952606}{10222818} = 0{,}97356776$

Como se puede ver en estos índices, la cantidad que tomamos como referencia es el número de visitantes de los parques del año anterior, por lo que podemos denominarlos «índices encadenados» y nos permitirá ver el crecimiento año tras año.

$I_{2000}^{2001} = 0{,}98$ nos indica que el número de visitantes ha disminuido un 2 % del año 2000 al 2001.

$I_{2001}^{2002} = 0{,}97$ nos indica que el número de visitantes ha disminuido un 3 % del año 2001 al 2002.

$I_{2002}^{2003} = 1{,}07$ nos indica que el número de visitantes ha disminuido un 7 % del año 2002 al 2003.

$I_{2003}^{2004} = 1{,}08$ nos indica que el número de visitantes ha disminuido un 8 % del año 2003 al 2004.

$I_{2004}^{2005} = 0{,}96$ nos indica que el número de visitantes ha disminuido un 4 % en 2004 a 2005.

$I_{2005}^{2006} = 1{,}02$ nos indica que el número de visitantes ha disminuido un 2 % en 2005 a 2006.

$I_{2006}^{2007} = 0{,}99$ nos indica que el número de visitantes ha disminuido un 1 % del año 2006 al 2007.

$I_{2007}^{2008} = 0{,}94$ nos indica que el número de visitantes ha disminuido un 6 % del año 2007 al 2008.

$I_{2008}^{2009} = 0{,}97$ nos indica que el número de visitantes ha disminuido un 3 % del año 2008 al 2009.

Estos índices dan idea del aumento del número total de visitantes en todo el período detallando la evolución por años, por lo que podemos diferenciar los años en que el número ha aumentado y en los que ha disminuido.

c) Calcula el incremento total e incremento medio anual de este período, a partir de las cantidades originales y a partir de los índices encadenados.

Para calcular el incremento total basándonos en las cantidades originales, tan solo hay que considerar el número de visitantes del primer y último año del período por analizar. Este concepto corresponde al último índice calculado en el apartado *a*). Así:

0,97 nos indica que si comparamos las visitas del año 2009 con las del año 2000, veremos que han disminuido un 3 %.

Para obtener el incremento medio anual, habrá que considerar la raíz con un índice 9 que viene dado por el número de incrementos que contiene el período estudiado. Notamos que corresponde al número de años o datos menos uno. Así, con los datos del ejercicio:

$$\sqrt[9]{\frac{9952606}{10252799}} = \sqrt[9]{0{,}97072087} = 0{,}9967$$

que podremos interpretar como que la evolución total de disminución del 3 % es equivalente a una disminución anual del 0,33 %. Diremos que el incremento anual medio es del –0,33 %.

Reseñamos que en estos cálculos que hemos hecho, tan solo hemos utilizado el número de visitantes de los parques de los años 2000 y 2009.

Veamos también cómo se pueden calcular estos mismos incrementos totales y medio, basándonos en los índices encadenados que hemos obtenido en el apartado *b*).

$$I_{2000}^{2009} = I_{2000}^{2001} \cdot I_{2001}^{2002} \cdot I_{2002}^{2003} \cdot I_{2003}^{2004} \cdot I_{2004}^{2005} \cdot I_{2005}^{2006} \cdot I_{2006}^{2007} \cdot I_{2007}^{2008} \cdot I_{2008}^{2009} =$$

$$= 0{,}97558891 \cdot 0{,}96590618 \cdot 1{,}06571334 \cdot 1{,}08143618 \cdot 0{,}96484919 \cdot 1{,}02196588 \cdot 0{,}98955032 \cdot 0{,}94091712 \cdot 0{,}97356776 = 0{,}97072087$$

Para conocer el incremento medio, calcularemos pues la raíz novena del producto de los índices (el índice de la raíz coincide con el número de índices que forman el producto).

$$\sqrt[9]{I_{2000}^{2001} \cdot I_{2001}^{2002} \cdot I_{2002}^{2003} \cdot I_{2003}^{2004} \cdot I_{2004}^{2005} \cdot I_{2005}^{2006} \cdot I_{2006}^{2007} \cdot I_{2007}^{2008} \cdot I_{2008}^{2009}} =$$

$$\sqrt[9]{0{,}9756 \cdot 0{,}9660 \cdot 1{,}0657 \cdot 1{,}0814 \cdot 0{,}9648 \cdot 1{,}0220 \cdot 0{,}9896 \cdot 0{,}9409 \cdot 0{,}9736} =$$

$$\sqrt[9]{0{,}9707} = 0{,}9967$$

que interpretaremos como un disminución anual constante del 0,33 %.

d) Haz previsiones del número de visitantes de los parques considerados para los años 2010, 2011 y 2012, si consideramos que no hubiera cambios significativos en el comportamiento de la afluencia.

Para hacer previsiones hay que partir de la hipótesis de que el incremento medio anual que hemos obtenido en el apartado anterior, podría ser una estimación del incremento anual de los años que están por venir y que, a partir del último dato conocido, calcularemos las cantidades de visitantes que podremos esperar para los años 2010, 2011 y 2012:

$9952606 \cdot 0{,}9967 = 9919762$ visitantes
$9952606 \cdot 0{,}9967^2 = 9887027$ visitantes
$9952606 \cdot 0{,}9967^3 = 9854400$ visitantes

Notamos que por previsiones o estimaciones partiremos del último dato real y lo multiplicaremos por el incremento medio anual que hemos calculado previamente, y elevándolo al número de años o períodos que están por venir.

Ejercicio 3

A continuación presentamos las variaciones porcentuales del volumen de ventas de cierta superficie comercial, en los últimos años.

Año	**Variaciones del volumen de ventas (%)**
2006	–3,13
2007	–2,15
2008	+2,12
2009	+3,15
2010	+4,12
2011	+4,31

a) Calcula los índices de las ventas de cada año, tomando como referencia el año 2005 y los índices encadenados.
b) Calcula la variación o incremento medio anual y total de las ventas en este período.
c) Estima las ventas de los dos años siguientes si suponemos que no hay cambios significativos en el comportamiento de las ventas en estos años.

Solución

a) Calcula los índices de las ventas de cada año, tomando como referencia el año 2005 y los índices encadenados.

Hay que ver que los datos de este ejercicio se diferencian de los dos anteriores, ya que en este caso los datos son incrementos porcentuales anuales, por lo que los datos de la tabla se podrán «traducir» y convertirse en índice encadenados, tal como se indica en la siguiente tabla:

Año	Variaciones del volumen de ventas (%)	Índices encadenados
2006	–3,13	$I_{2005}^{2006} = 0{,}9687$
2007	–2,15	$I_{2006}^{2007} = 0{,}9785$
2008	+2,12	$I_{2007}^{2008} = 1{,}0212$
2009	+3,15	$I_{2008}^{2009} = 1{,}0315$
2010	+4,12	$I_{2009}^{2010} = 1{,}0412$
2011	+4,31	$I_{2010}^{2011} = 1{,}0431$

Podemos ver cómo calculamos estos índices. A tal fin, le sumamos o restamos a 1 el incremento, y así convertiremos el tanto por ciento en tanto por uno:

$I_{2005}^{2006} = 1 - 0.0313 = 0{,}9687$

$I_{2006}^{2007} = 1 - 0{,}0215 = 0{,}9785$

$I_{2007}^{2008} = 1 + 0{,}0212 = 1{,}0212$

$I_{2008}^{2009} = 1 + 0{,}0315 = 1{,}0315$

Y así con el resto de valores de la columna de la derecha de la tabla.

Para calcular los índices de base 2005, multiplicaremos los índices anteriores:

$I_{2005}^{2007} = I_{2005}^{2006} \cdot I_{2006}^{2007} = 0{,}9687 \cdot 0{,}9785 = 0{,}9479$

$I_{2005}^{2008} = I_{2005}^{2006} \cdot I_{2006}^{2007} \cdot I_{2007}^{2008} = 0{,}9687 \cdot 0{,}9785 \cdot 1{,}0212 = 0{,}9680$

$I_{2005}^{2009} = I_{2005}^{2006} \cdot I_{2006}^{2007} \cdot I_{2007}^{2008} \cdot I_{2008}^{2009} = 0{,}9687 \cdot 0{,}9785 \cdot 1{,}0212 \cdot 1{,}0315 = 0{,}9985$

$I_{2005}^{2010} = I_{2005}^{2006} \cdot I_{2006}^{2007} \cdot I_{2007}^{2008} \cdot I_{2008}^{2009} \cdot I_{2009}^{2010} =$

$= 0{,}9687 \cdot 0{,}9785 \cdot 1{,}0212 \cdot 1{,}0315 \cdot 1{,}0412 = 1{,}0396$

$I_{2005}^{2011} = I_{2005}^{2006} \cdot I_{2006}^{2007} \cdot I_{2007}^{2008} \cdot I_{2008}^{2009} \cdot I_{2009}^{2010} \cdot I_{2010}^{2011} =$

$= 0{,}9687 \cdot 0{,}9785 \cdot 1{,}0212 \cdot 1{,}0315 \cdot 1{,}0412 \cdot 1{,}0431 = 1{,}0844$

Presentaremos todos los índices en la siguiente tabla:

Año	Variaciones del volumen de ventas (%)	Índice encadenados	Índice Base 2005
2006	–3,13	$I_{2005}^{2006} = 0{,}9687$	$I_{2005}^{2006} = 0{,}9687$
2007	–2,15	$I_{2006}^{2007} = 0{,}9785$	$I_{2005}^{2007} = 0{,}9479$
2008	+2,12	$I_{2007}^{2008} = 1{,}0212$	$I_{2005}^{2008} = 0{,}9680$
2009	+3,15	$I_{2008}^{2009} = 1{,}0315$	$I_{2005}^{2009} = 0{,}9985$
2010	+4,12	$I_{2009}^{2010} = 1{,}0412$	$I_{2005}^{2010} = 1{,}0396$
2011	+4,31	$I_{2010}^{2011} = 1{,}0431$	$I_{2005}^{2011} = 1{,}0844$

Si interpretamos los índices de la última columna podremos ver el incremento global en el volumen de ventas de la superficie, en la casilla inferior. Este índice indica un aumento de +8,44 % en el período analizado.

b) Calcula la variación o incremento medio anual y total de las ventas en este período.

Para calcular el incremento total de las ventas tan solo hay que interpretar el índice $I_{2005}^{2011} = 1.0844$, el cual nos indica que las ventas han aumentado un 8,44 % en el período analizado.

Podemos recordar que este índice lo hemos calculado multiplicando los índices encadenados que hemos «construido» con los datos de los porcentajes del enunciado.

Para determinar el incremento medio anual, calculamos la raíz de índice 6, ya que hacemos una media geométrica con los índices de cada año.También podemos partir del índice que representa el incremento total del período.

Así, el incremento medio anual es

$$\sqrt[6]{0{,}9687 \cdot 0{,}9785 \cdot 1{,}0212 \cdot 1{,}0315 \cdot 1{,}0412 \cdot 1{,}0431} = \sqrt[6]{1{,}0844} = 1{,}0136$$

que podemos interpretar como que el aumento total es equivalente a un incremento anual constante del 1,36 %.

c) Estima las ventas de los dos años siguientes si suponemos que no hay cambios significativos en el comportamiento de las ventas en estos años.

No podemos estimar las ventas porque no conocemos la magnitud de las ventas de ningún año para cogerlo de referencia ya partir de él calcular las ventas del año que nos interesa.

Ejercicio 4

A continuación presentamos los valores del índice de precios al consumo, IPC, que podemos consultar en la página del INE y que hace referencia a los datos en base a 2001 y 2006.

Por razones que habrá que estudiar en la teoría, en ciertos momentos hay que hacer un cambio en el año de referencia y se empieza a obtener la nueva serie del IPC, comenzado de nuevo con el valor 100. Diremos que ha habido un «cambio de base».

A menudo, como podrás ver en ejercicios posteriores, hay que utilizar en un mismo cálculo el valor del IPC de años que corresponden a períodos de bases diferentes, y necesitaremos trabajar con todos los valores del IPC referidos a una misma base. Estos datos los podrás encontrar fácilmente en la página web del INE, pero en este ejercicio vamos a ver cómo se calculan los valores de las casillas que están sombreadas en gris.

En primer lugar, presentamos la tabla de los valores del IPC desde el año 2002 al 2006 en base 2001.

Índice de precios al consumo	
Medias anuales. Base 2001	
Nacional por general y Grupos COICOP	
Unidades: Índice y tasas	
	General
	Media anual
2006	117,624
2005	113,63
2004	109,927
2003	106,684
2002	103,538

Y a continuación, los datos de los valores del IPC desde el año 2006 al 2010 en base 2006, aunque están añadidos los valores de las casillas gris que corresponden a los valores obtenidos «a posteriori» para facilitar los trabajos de cálculo referidos a períodos de diferentes bases.

Índice de precios al consumo	
Medias anulales. Base 2006	
Índices nacionales: general y de grupos COICOP	
Unidades: Base 2006=100	
	General
	Media anual
2010	108,588
2009	106,668
2008	106,976
2007	102,787
2006	100
2005	96,604
2004	93,456
2003	90,699
2002	88,024

Explica cómo se han obtenido los datos de las casillas sombreadas en gris, averiguando el valor del enlace.

Fuente: INE

Solución

Este proceso que denominamos «cambio de base» tan solo es la transformación de los valores del IPC para asegurarnos la proporcionalidad en la cadena de los valores de los períodos anteriores al momento en que, por razones que no vienen al caso, el INE realiza esta actualización y, por tanto, comienza una nueva serie de valores partiendo del 100.

Veamos que el cálculo es tan solo la resolución de una proporcionalidad (regla de tres) donde hay términos fijos que nos permitirán encontrar el valor del enlace.

En la primera tabla tenemos los valores del IPC correspondientes al período 2002-2006 en base 2001. Pero llegado el año 2006 se decide un cambio de base y como se puede ver en la segunda tabla, tenemos una nueva serie de datos que comienza en el año 2006 con un 100. En alguna situación que veremos en ejercicios posteriores necesitamos utilizar el IPC de los dos períodos, por lo que es necesario conocer todos los datos referidos a la misma base para no romper la continuidad de la secuencia que refleja el comportamiento de los precios al consumo y, por consiguiente, se convierte en uno de los principales indicadores de la inflación y de los devenires económicos de un país.

Veamos, pues, cómo se calculan los valores de las casillas sombreadas en gris. Son los valores de los antiguos IPC en la nueva base. Para su cálculo, tan solo hay que plantear los datos implicados para asegurarnos la proporcionalidad real en la evolución de los precios.

Veamos los datos de 2002:

Para calcular el dato desconocido

$$IPC_{2006}^{2002} = \frac{IPC_{2001}^{2002} \cdot IPC_{2006}^{2006}}{IPC_{2001}^{2006}} = \frac{103{,}538 \cdot 100}{117{,}624} = 103{,}538 \frac{100}{117{,}624} = 103{,}538 \cdot 0{,}8502 =$$
$$= 88{,}025$$

Del mismo modo, el resto de años repetimos el proceso:

$$IPC_{2006}^{2003} = \frac{IPC_{2001}^{2003} \cdot IPC_{2006}^{2006}}{IPC_{2001}^{2006}} = \frac{106{,}684 \cdot 100}{117{,}624} = 106{,}684 \frac{100}{117{,}624} = 106{,}684 \cdot 0{,}8502 =$$
$$= 90{,}699$$

$$IPC_{2006}^{2004} = \frac{IPC_{2001}^{2004} \cdot IPC_{2006}^{2006}}{IPC_{2001}^{2006}} = \frac{109{,}927 \cdot 100}{117{,}624} = 109{,}927 \frac{100}{117{,}624} = 109{,}927 \cdot 0{,}8502 =$$
$$= 93{,}456$$

$$IPC_{2006}^{2005} = \frac{IPC_{2001}^{2005} \cdot IPC_{2006}^{2006}}{IPC_{2001}^{2006}} = \frac{113{,}63 \cdot 100}{117{,}624} = 113{,}63 \frac{100}{117{,}624} = 113{,}63 \cdot 0{,}8502 =$$

96,604

Es evidente que para obtener los valores del IPC del año «y» en base 2006 hemos multiplicado el IPC del año «y» en base 2001 por la fracción $\frac{IPC_{2006}^{2006}}{IPC_{2001}^{2006}}$ el valor de la que hemos reseñado en rojo y es lo que denominaremos «enlace técnico».

Aunque no se contempla en el ejercicio, si dividimos los valores del IPC de la segunda tabla por este «enlace técnico», también podríamos obtener los IPC del período 2006-2010 en base 2001.

La tabla completa quedaría:

	2002	2003	2004	2005	2006	2007	2008	2009	2010
IPC Base2001	103,538	106,684	109,927	113,63	117,624	120,902	125,829	125,467	127,726
IPC Base2006	88,024	90,699	93,456	96,604	100	102,787	106,976	106,668	108,588

Ejercicio 5

Calcula los índices de precios y cantidades de los artículos A, B y C mediante las fórmulas de Laspeyres y Paasche, de los años 2008, 2009 y 2010 en función del año 2008, utilizando los datos de las siguientes tablas donde están indicadas las cantidades q_i y precios p_i que hay que conocer.

	2008		**2009**		**2010**	
	Precio p_i	Cantidad q_i	Precio p_i	Cantidad q_i	Precio p_i	Cantidad q_i
Art. A	12	100	14	112	15	115
Art. B	10	50	8	65	7	72
Art. C	5	20	10	10	15	5

Solución

Índice de precios de Laspeyres

Para calcular estos índices, empezaremos por conocer y deducir la fórmula que emplearemos. Hemos reducido su cálculo a tres artículos pero no olvidemos que este cálculo se extiende a la totalidad de artículos representativos del consumo de las familias en un país (véase ECPF).

$$L_{precios}{}_0^t = \frac{\sum_i \frac{p_{it}}{p_{i0}} \cdot p_{i0} \cdot q_{i0}}{\sum_i p_{i0} \cdot q_{i0}} = \frac{\sum_i p_{it} \cdot q_{i0}}{\sum_i p_{i0} \cdot q_{i0}}$$

donde *t* es el año actual y *0* será el año que tomaremos como referencia en la comparación. Si se trata de índice encadenados podríamos decir años *t-1* y *t*.

p_{it} será el precio del artículo y el año *t*
p_{i0} será el precio del artículo y el año *0*
q_{it} será la cantidad del artículo y el año *t*
q_{i0} será la cantidad del artículo y el año *0*

es una media ponderada donde el «peso» de cada artículo $p_{i0} \cdot q_{i0}$ es el valor del artículo en la «cesta de la compra» del año de referencia y permanecerá constante a lo largo del período mientras no se cambie la base. Un inconveniente de este método es que si la importancia de los artículos en los hábitos de consumo cambia mucho, estos coeficientes quedan desfasados.

Así:

$$L_{precios}{}_{2008}^{2008} = 1$$

$$L_{precios}{}_{2008}^{2009} = \frac{\sum_i p_{i2009} \cdot q_{i2008}}{\sum_i p_{i2008} \cdot q_{i2008}} = \frac{14 \cdot 100 + 8 \cdot 50 + 10 \cdot 20}{12 \cdot 100 + 10 \cdot 50 + 5 \cdot 20} = \frac{2000}{1800} = 1{,}111$$

$$L_{precios}{}_{2008}^{2010} = \frac{\sum_i p_{i2010} \cdot q_{i2008}}{\sum_i p_{i2008} \cdot q_{i2008}} = \frac{15 \cdot 100 + 7 \cdot 50 + 15 \cdot 20}{12 \cdot 100 + 10 \cdot 50 + 5 \cdot 20} = \frac{2150}{1800} = 1{,}194$$

Se puede comprobar que el denominador no varía y tan solo hay que actualizar los precios de los artículos en el período nuevo por comparar. Esto es una gran ventaja de esta fórmula.

Índice de cantidades de Laspeyres

En este caso vamos a estudiar la evolución de las cantidades demandadas y para la ponderación se utilizan los mismos coeficientes del apartado anterior $p_{i0} \cdot q_{i0}$.

$$L_{cantidades}{}_0^t = \frac{\sum_i \frac{q_{it}}{q_{i0}} \cdot p_{i0} \cdot q_{i0}}{\sum_i p_{i0} \cdot q_{i0}} = \frac{\sum_i q_{it} \cdot p_{i0}}{\sum_i p_{i0} \cdot q_{i0}}$$

Así:

$$L_{cantidades\,2008}^{\;\;2008} = 1$$

$$L_{cantidades\,2008}^{\;\;2009} = \frac{\sum_i q_{i2009} \cdot p_{i2008}}{\sum_i p_{i2008} \cdot q_{i2008}} = \frac{112 \cdot 12 + 65 \cdot 10 + 10 \cdot 5}{12 \cdot 100 + 10 \cdot 50 + 5 \cdot 20} = \frac{2044}{1800} = 1{,}136$$

$$L_{cantidades\,2008}^{\;\;2010} = \frac{\sum_i p_{i2010} \cdot q_{i2008}}{\sum_i p_{i2008} \cdot q_{i2008}} = \frac{115 \cdot 12 + 72 \cdot 10 + 50 \cdot 5}{12 \cdot 100 + 10 \cdot 50 + 5 \cdot 20} = \frac{2350}{1800} = 1{,}306$$

Índice de precios de Paasche

$$P_{precios\,0}^{\;\;t} = \frac{\sum_i \frac{p_{it}}{p_{i0}} \cdot p_{i0} \cdot q_{it}}{\sum_i p_{i0} \cdot q_{it}} = \frac{\sum_i p_{it} \cdot q_{it}}{\sum_i p_{i0} \cdot q_{it}}$$

es una media ponderada donde el «peso» de cada artículo $p_{i0} \cdot q_{it}$ intenta mejorar la propuesta de Laspeyres, evitando en cierto modo el desfase, ya que recoge la importancia del artículo al considerar la cantidad en el período comparar.

Así:

$$L_{precios\,2008}^{\;\;2008} = 1$$

$$P_{precios\,2008}^{\;\;2009} = \frac{\sum_i p_{i2009} \cdot q_{i2009}}{\sum_i p_{i2008} \cdot q_{i2009}} = \frac{14 \cdot 112 + 8 \cdot 65 + 10 \cdot 10}{12 \cdot 112 + 10 \cdot 65 + 5 \cdot 10} = \frac{2188}{2044} = 1{,}070$$

$$P_{precios\,2008}^{\;\;2010} = \frac{\sum_i p_{i2010} \cdot q_{i2010}}{\sum_i p_{i2008} \cdot q_{i2010}} = \frac{15 \cdot 115 + 7 \cdot 72 + 15 \cdot 50}{12 \cdot 115 + 10 \cdot 72 + 5 \cdot 50} = \frac{2979}{2350} = 1{,}268$$

Índice de catidades de Paasche

$$P_{cantidades}{}_0^t = \frac{\sum_i \frac{q_{it}}{q_{i0}} \cdot q_{i0} \cdot p_{it}}{\sum_i q_{i0} \cdot p_{it}} = \frac{\sum_i p_{it} \cdot q_{it}}{\sum_i q_{i0} \cdot p_{it}}$$

Esta propuesta, como que analiza la evolución de las cantidades, considera como coeficiente $q_{i0} \cdot p_{it}$ que indica el «peso» de cada artículo, el precio del año t para actualizar la importancia del artículo.

Así:

$$L_{cantidades}{}_{2008}^{2008} = 1$$

$$P_{cantidades}{}_{2008}^{2009} = \frac{\sum_i p_{i2009} \cdot q_{i2009}}{\sum_i q_{i2008} \cdot p_{i2009}} = \frac{14 \cdot 112 + 8 \cdot 65 + 10 \cdot 10}{14 \cdot 100 + 8 \cdot 50 + 10 \cdot 20} = \frac{2188}{2000} = 1{,}094$$

$$P_{cantidades}{}_{2008}^{2010} = \frac{\sum_i p_{i2010} \cdot q_{i2010}}{\sum_i q_{i2008} \cdot p_{i2010}} = \frac{15 \cdot 115 + 7 \cdot 72 + 15 \cdot 50}{15 \cdot 100 + 7 \cdot 50 + 15 \cdot 20} = \frac{2979}{2150} = 1{,}386$$

Como se puede ver en estos índices, en cada uno calculado por las fórmulas de Paasche, hay que determinar siempre tanto el numerador como el denominador de cada fracción. Esta diferencia que nos puede parecer irrelevante para tres artículos, no lo parece igual para la gran cantidad de datos que hay que trabajar para el cálculo del IPC y con los recursos tecnológicos de tiempo atrás.

Ejercicio 6

Supongamos que compramos una vivienda por 16.125.000 ptas. en diciembre de 1998 y la hemos vendido en diciembre de 2006 por un valor de 240.000 euros. Averigua el porcentaje de beneficios o pérdidas que hemos tenido en la operación.

Nota: Para realizar las operaciones consultaremos los valores del IPC que necesitamos en la página web del INE. www.ine.es (sería interesante calcular este incremento con el IPC general y con el IPC del grupo de la vivienda).

El cambio de moneda que consideraremos es 1 € = 166,386 ptas.

Solución

Para empezar a comparar habrá que trabajar en una única moneda. Decidimos trabajar en euros. Es obvio que el resultado en términos relativos o porcentajes no varía si trabajamos en pesetas.

Para transformar 16.125.000 ptas. a euros utilizaremos el cambio que propone la nota (1 € = 166,386 ptas.) por lo que, 16.125.000 / 166,386 = 96.913,20 euros en términos corrientes de diciembre de 1998. Para averiguar cuál sería su valor equivalente en términos corrientes del año 2006, hay que hacer la siguiente transformación:

$$96913{,}20 \cdot \frac{IPC_{2001}^{2006}}{IPC_{2001}^{1998}} = 96913{,}20 \cdot \frac{117{,}624}{91{,}223} = 124961{,}01 \text{ €}$$

↓

En las páginas del INE se pueden obtener estos datos.

$$IPC_{1992}^{1998} = 123{,}791 \qquad IPC_{2001}^{2001} = 100$$
$$IPC_{1992}^{2001} = 135{,}702 \qquad IPC_{2001}^{2006} = 117{,}624$$

Y como necesitaban averiguar el valor de IPC_{2001}^{1998}, hemos procedido como se explica en el ejercicio 4:

$$IPC_{2001}^{1998} = \frac{IPC_{1992}^{1998} \cdot IPC_{2001}^{2001}}{IPC_{1992}^{2001}} = \frac{123{,}791 \cdot 100}{135{,}702} = 91{,}223$$

Estos 124.961,01 euros serían el valor equivalente, en cuanto a poder adquisitivo, del valor de compra de la vivienda en el año 2006.

Como en el enunciado se dice que la hemos vendido por 240.000 euros, vamos a calcular los beneficios en términos relativos a partir del concepto de índice:

$$\frac{240000}{124961{,}01} = 1{,}92$$

Este cociente nos permite interpretar que tenemos un beneficio del 92 %, es decir, casi se ha duplicado el valor de la vivienda en el período de 8 años que hemos contemplado.

Nota: Para realizar los cálculos hemos utilizado las medias anuales del IPC, pero se podría hacer también con los valores del IPC exactamente los meses de compra y venta, así como elegir los IPC del grupo de vivienda en lugar de la IPC general. Dejemos estas variantes para el trabajo del lector.

Ejercicio 7

En la siguiente tabla mostramos los datos de los impuestos municipales de cierta vivienda en los últimos años.

Año	*Importe impuesto municipal (términos nominales)*
2006	503,24
2007	515,65
2008	536,73
2009	578,84
2010	584,42

Para analizar su evolución,

a) Deflacta la serie, convirtiéndola en monedas constantes del 2006.
b) Calcula los índices que nos permitirán estudiar su evolución año por año, en términos reales o monedas constantes del año 2006. Interpreta los resultados.
c) Calcula el incremento total e incremento medio en el período en términos reales.
d) Suponiendo que los impuestos sigan este comportamiento, averigua el valor en términos nominales o monedas corrientes para los años 2011, 2012 y 2013.

Nota: Para resolver este ejercicio, utilizaremos los valores de la media anual del IPC general que necesitamos, obteniéndose los de la página web del INE. www.ine.es.

Solución

Para analizar su evolución,

a) Deflacta la serie, convirtiéndola en monedas constantes del 2006.

A tal fin, es necesario que consultemos los valores del IPC de estos años. Tomaremos las medias anuales del IPC general. Hay que insistir en que todos los índices que trabajamos en el mismo ejercicio deben estar en la misma base; de lo contrario, hay que hacer el cambio de base que proceda, tal y como se explicó en el ejercicio 4.

Presentamos los resultados en la siguiente tabla:

Año	Importe impuesto municipal (términos nominales)	Importe impuesto municipal (términos reales 2006)
2006	503,24	503,24
2007	515,65	$515{,}65 \cdot \frac{IPC_{2006}^{2006}}{IPC_{2006}^{2007}} = 515{,}65 \cdot \frac{100}{102{,}787} = 501{,}67$
2008	536,73	$536{,}73 \cdot \frac{IPC_{2006}^{2006}}{IPC_{2006}^{2008}} = 536{,}73 \cdot \frac{100}{106{,}976} = 501{,}73$
2009	578,84	$578{,}84 \cdot \frac{IPC_{2006}^{2006}}{IPC_{2006}^{2009}} = 578{,}84 \cdot \frac{100}{106{,}668} = 542{,}66$
2010	584,42	$584{,}42 \cdot \frac{IPC_{2006}^{2006}}{IPC_{2006}^{2010}} = 584{,}42 \cdot \frac{100}{108{,}588} = 538{,}20$

Con esta operación, le hemos «eliminado» al importe del impuesto, el efecto de la inflación y podremos analizar «en términos reales» su evolución como tal magnitud, salvo las influencias de los devenires de la economía general que se reflejan en las variaciones del índice de precios.

b) Calcula los índices que nos permitirán estudiar su evolución año por año, en términos reales o monedas constantes del año 2006. Interpreta los resultados.

Nos piden los índices encadenados con los valores de la última columna de la tabla anterior:

Año	Importe impuesto municipal (términos nominales)	Importe impuesto municipal (términos reales 2006)	Índice	Interpretación
2006	503,24	503,24	-----	
2007	515,65	501,67	$I_{2006}^{2007} = \frac{501{,}67}{503{,}24} = 0{,}997$	Ha disminuido un 0,3 %
2008	536,73	501,73	$I_{2007}^{2008} = \frac{501{,}73}{501{,}67} = 1{,}0001$	Podemos considerar que es constante
2009	578,84	542,66	$I_{2008}^{2009} = \frac{542{,}66}{501{,}73} = 1{,}082$	Ha aumentado un 8,2 %
2010	584,42	538,20	$I_{2009}^{2010} = \frac{538{,}20}{542{,}66} = 0{,}992$	Ha disminuido un 0,8 %

En general, vemos que en términos reales era un importe que permanece estable en el período analizado, ya que la evolución del importe es paralela a la evolución del IPC, excepto en el año 2009 que de manera puntual hace un aumento del 8,2 %.

Podemos ver con más claridad esta evolución, cuando hemos «borrado» el efecto de la inflación.

c) Calcula el incremento total e incremento medio en el período en términos reales.

Para calcular los incrementos que nos planteamos, es más cómodo partir de los datos de la magnitud. En este caso, nos referimos al importe del impuesto municipal en términos reales del 2006.

Para calcular el incremento total del período, interpretaremos el índice:

$$I_{2006}^{2010} = \frac{538,20}{503,24} = 1,069$$

que nos permite afirmar que el importe del impuesto ha aumentado un 6,9 % en términos reales en el período considerado.

Para calcular el incremento medio anual, calcularemos la raíz siguiente:

$$\sqrt[4]{1,069} = 1,017$$

que nos permite afirmar que el incremento total del 6,9 % es equivalente a un incremento constante anual del 1,7 % durante 4 años.

Queremos señalar que estos resultados también se podrían obtener a partir de los índices «encadenados», aunque no es razonable si disponemos de los valores de la magnitud por analizar.

Así, el incremento total del período sería:

$$0,997 \cdot 1,0001 \cdot 1,082 \cdot 0,992 = 1,07$$

que sería un 7 % de aumento total en el período. La diferencia (un décima) se debe a los errores del redondeo de cada uno de los índices.

El incremento medio anual se obtendría también:

$$\sqrt[4]{0,997 \cdot 1,0001 \cdot 1,082 \cdot 0,992} = \sqrt[4]{1,0702} = 1,017$$

que da el mismo resultado que hemos comentado antes.

d) Suponiendo que los impuestos sigan este comportamiento, averigua el valor en términos nominales o monedas corrientes para los años 2011, 2012 y 2013.

Como nos piden que demos el resultado en moneda corriente o términos nominales tendremos que estimar los posibles valores del IPC en los años venideros, para operar de una manera similar en el apartado *a*) pero en sentido opuesto.

A tal fin, obtendremos el incremento medio anual del IPC de los años del período estudiado, a partir de los valores del primer y último año.

$\frac{IPC_{2006}^{2010}}{IPC_{2006}^{2006}} = \frac{108,588}{100} = 1,08588$ y para obtener el incremento medio anual del IPC del período calcularemos la raíz cuarta correspondiente,

$\sqrt[4]{1,08588} = 1,0208$ que nos permite afirmar que dicho incremento es del 2,08 % anual.
Así, basándonos en este resultado, podremos estimar el IPC de los siguientes años:

$$IPC_{2006}^{2011} = IPC_{2006}^{2010} \cdot 1,0208 = 108,588 \cdot 1,0208 = 110,847$$

$$IPC_{2006}^{2012} = IPC_{2006}^{2010} \cdot 1,0208^2 = 108,588 \cdot 1,0208^2 = 113,152$$

$$IPC_{2006}^{2013} = IPC_{2006}^{2010} \cdot 1,0208^3 = 108,588 \cdot 1,0208^3 = 115,506$$

Calcularemos primero el importe del impuesto en términos reales, aplicando el incremento medio anual (1,7 %) sobre el valor de este importe del año 2010, es decir, 538,20 euros en términos reales de 2006.

Año 2011 → $538,20 \cdot 1,017 = 547,25$
Año 2012 → $538,20 \cdot 1,017^2 = 556,46$
Año 2013 → $538,20 \cdot 1,017^3 = 565,82$

Como estos resultados están expresados en términos reales del año 2006, hay que convertirlos a términos nominales:

Año 2011 → $547,25 \cdot \frac{IPC_{2006}^{2011}}{IPC_{2006}^{2006}} = 547,25 \cdot \frac{110,847}{100} = 606,61$

Año 2012 → $556,46 \cdot \frac{IPC_{2006}^{2012}}{IPC_{2006}^{2006}} = 556,46 \cdot \frac{113,152}{100} = 629,65$

Año 2013 → $565,82 \cdot \frac{IPC_{2006}^{2013}}{IPC_{2006}^{2006}} = 565,82 \cdot \frac{115,506}{100} = 653,56$

Tan solo notar que esta estimación está hecha bajo la hipótesis de que tanto el IPC como el importe del impuesto, evolucionará al ritmo anual que indique el incremento medio anual de cada una de las magnitudes.

En este ejercicio, a pesar de todo, parece que este incremento no refleja la realidad del comportamiento del importe del impuesto, que ya hemos comentado que ha sido estable la mayor parte del período y solo experimentó un importante aumento del 8 % en el año 2009, de manera puntual. Esta matización hace que las estimaciones, en cierto modo, pierdan cierta fiabilidad.

Ejercicio 8

En la tabla siguiente se indica el valor de la nómina mensual de un trabajador en los últimos años.

Año	Nómina mensual (€)
2007	2034,75
2008	2062,13
2009	2218,61
2010	2253,67
2011	2181,75

Estudia la pérdida o ganancia de su poder adquisitivo para cada año y de todo el período global, considerando los valores de la media anual del IPC general que puedes encontrar en la página del INE.

Solución

Sería conveniente recordar en este momento el concepto de pérdida o ganancia de poder adquisitivo. Se puede decir que nosotros ganamos poder adquisitivo (capacidad de compra de bienes de consumo) si el salario que percibimos este año está por encima de lo percibiríamos si nuestro salario hubiera sido incrementado en el mismo porcentaje que aumentan los precios de estos bienes. Podríamos razonar de la misma manera para definir la pérdida de poder adquisitivo cuando nuestro salario queda por debajo de lo que tendríamos si la hubieran incrementado con el mismo porcentaje que los precios.

El incremento de estos precios está reflejado en el IPC que publica el INE cada mes. Nosotros tomaremos la media anual general de este índice que podremos encontrar fácilmente en la web de este organismo.

Ahora bien, como hacemos un análisis en términos relativos y damos el resultado en porcentajes, veamos en la siguiente expresión, cómo el salario concreto del que partimos, no es necesario en el estudio de la evolución del poder adquisitivo:

Ganancia o pérdida de poder adquisitivo = $\Delta_{poder\ adquisitivo} =$

$$= \frac{salario\ nuevo}{salario\ actualizado\ según\ IPC} = \frac{salario\ anterior \cdot (1+\Delta_{salarial})}{salario\ anterior \cdot (1+\Delta_{IPC})} = \frac{(1+\Delta_{salarial})}{(1+\Delta_{IPC})}$$

La pérdida o ganancia del poder adquisitivo, pues, se calcula a partir de la comparación de los incrementos anuales del salario ($\Delta_{salario}$) y del IPC (Δ_{IPC},) paralelamente.

A tal fin, comenzaremos por calcular los índices «encadenados» de los salarios, que nos permitirán averiguar los incrementos salariales anuales. En la siguiente tabla se detallan cálculos y resultados.

Año	*Nómina mensual (€)*	Índice	
2007	2034,75	------	-------
2008	2062,13	$I_{2007}^{2008} = \frac{2062,13}{2034,75} = 1,014$	Ha aumentado 1,4 %
2009	2218,61	$I_{2008}^{2009} = \frac{2218,61}{2062,13} = 1,076$	Ha aumentado 7,6 %
2010	2253,67	$I_{2009}^{2010} = \frac{2253,67}{2218,61} = 1,016$	Ha aumentado 1,6 %
2011	2181,75	$I_{2010}^{2011} = \frac{2181,75}{2253,67} = 0,968$	Ha disminuido 3,2 %

Consultaremos la página del INE para encontrar los valores del IPC de estos años. Nos interesa la media anual del índice general en base 2006. Si no se dispone de los datos del año 2011, se utiliza la estimación que se obtiene en el ejercicio 7 de esta colección. Y con estos datos haremos los mismos análisis que hemos hecho con las nóminas para obtener los incrementos anuales.

Año	*IPC Base 2006*	Índice	
2007	$IPC_{2006}^{2007} = 102,787$	------	-------
2008	$IPC_{2006}^{2008} = 106,976$	$I_{2007}^{2008} = \frac{106,976}{102,787} = 1,041$	Ha aumentado 4,1 %
2009	$IPC_{2006}^{2009} = 106,668$	$I_{2008}^{2009} = \frac{106,668}{106,976} = 0,997$	Ha disminuido 0,3 %
2010	$IPC_{2006}^{2010} = 108,588$	$I_{2009}^{2010} = \frac{108,588}{106,668} = 1,018$	Ha aumentado 1,8 %
2011	$IPC_{2006}^{2011} = 110,847$	$I_{2010}^{2011} = \frac{110,847}{108,588} = 1,021$	Ha aumentado 2,1 %

Ya hemos explicado al comenzar el ejercicio el concepto de *pérdida o ganancia de poder adquisitivo*. Se gana poder adquisitivo cuando el incremento salarial está por encima del incremento del IPC que nos indica, asimismo, el incremento de los precios de los bienes de consumo. Del mismo modo, habrá una pérdida de poder adquisitivo cuando el incremento salarial esté por debajo del incremento del IPC.

Para hacer esta comparación, partiremos de los índices que hemos calculado en las dos tablas anteriores y, como se hace un estudio en términos relativos, haremos los cocientes de estas cantidades año por año. Mostramos los resultados en la siguiente tabla y los cálculos de las casillas sombreadas en gris están debajo de la tabla.

	2008	2009	2010	2011
Incremento salarial	1,014	1,076	1,016	0,968
Incremento IPC	1,041	0,997	1,018	1,021
Pérdida o ganancia poder adquisitivo	0,974	1,079	0,998	0,948

Año 2008 → $\frac{1{,}014}{1{,}041} = 0{,}974$ Año 2009 → $\frac{1{,}076}{0{,}997} = 1{,}079$

Año 2010 → $\frac{1{,}016}{1{,}018} = 0{,}998$ Año 2011 → $\frac{0{,}968}{1{,}021} = 0{,}948$

Queda más claro si anotamos las interpretaciones en porcentajes, y así lo mostramos en la siguiente tabla. Convendremos que el signo positivo indica ganancia de poder adquisitivo y el signo negativo, pérdida.

	2008	2009	2010	2011
Incremento salarial	+1,4 %	+7,6 %	+1,6 %	–3,2 %
Incremento IPC	+4,1 %	–0,3 %	+1,8 %	+2,1 %
Pérdida o ganancia poder adquisitivo	–2,6 %	+7,9 %	–0,2 %	–5,2 %

Si queremos analizar el incremento total de los tres conceptos, podemos utilizar las magnitudes originales que disponemos tanto en lo que respecta a los salarios como al IPC, y lo haremos a partir de sus índices, para el poder adquisitivo.

Veamos el incremento salarial del total del período:

$$I_{salarios}{}_{2007}^{2011} = \frac{2181{,}75}{2034{,}75} = 1{,}072 \rightarrow \text{Ha aumentado un } 7{,}2\ \%$$

Calculemos ahora el incremento del IPC (recordemos que la fecha del IPC del 2011 es una estimación).

$$I_{IPC}{}_{2007}^{2011} = \frac{IPC_{2006}^{2011}}{IPC_{2006}^{2007}} = \frac{110{,}847}{102{,}787} = 1{,}078 \rightarrow \text{Ha aumentado un 7,8 \%}$$

Para calcular la pérdida de poder adquisitivo, planteamos el cociente de estos incrementos en su expresión de índice.

$$I_{poder\ adquisitiu}{}_{2007}^{2011} = \frac{I_{salaris}{}_{2007}^{2011}}{I_{IPC}{}_{2007}^{2011}} = \frac{1{,}072}{1{,}078} = 0{,}994 \rightarrow \text{Ha disminuido un 0,6 \%}$$

Si completamos la tabla anterior con esta información tenemos detallada la evolución total.

	2008	2009	2010	2011	TOTAL
Incremento salarial	+1,4 %	+7,6 %	+1,6 %	–3,2 %	+7,2 %
Incremento IPC	+4,1 %	–0,3 %	+1,8 %	+2,1 %	+7,8 %
Pérdida o ganancia poder adquisitivo	–2,6 %	+7,9 %	–0,2 %	–5,2 %	–0,6 %

Advertimos que aunque de un vistazo nos pueda parecer que los resultados de las casillas sombreadas se podrían obtener sumando y restando los porcentajes en filas y columnas, hay que fijarse en que no es cierto tal y como se puede comprobar con las datos totales y en algunas columnas, por ejemplo en el año 2008.

Ahora bien, sí podemos obtener los resultados, a partir de los índices de la tabla previa, multiplicándolos.

Veamos el incremento salarial del total del período:

$$I_{salaris}{}_{2007}^{2011} = 1{,}014 \cdot 1{,}076 \cdot 1{,}016 \cdot 0{,}968 = 1{,}073 \rightarrow \text{Ha aumentado un 7,3 \%}$$

Calculemos ahora el incremento del IPC.

$$I_{IPC}{}_{2007}^{2011} = 1{,}041 \cdot 0{,}997 \cdot 1{,}018 \cdot 1{,}021 = 1{,}079 \rightarrow \text{Ha aumentado un 7,9 \%}$$

Para calcular la pérdida de poder adquisitivo, planteamos también el producto de los factores:

$$I_{poder\ adquisitiu}{}_{2007}^{2011} = 0{,}974 \cdot 1{,}079 \cdot 0{,}998 \cdot 0{,}948 = 0{,}994 \rightarrow \text{Ha disminuido un 0,6 \%}$$

Hay que advertir que la diferencia con los resultados anteriores (del orden de décimas) se debe al redondeo de cada factor. Por esta razón insistimos en la recomendación de utilizar los datos originales de las magnitudes por analizar si disponemos de estas, pero presentamos los dos métodos de resoluciones, para los casos en que la información disponible sean los incrementos porcentuales anuales.

Ejercicio 9

En las tablas siguientes se presentan los valores del IPC y el incremento salarial de un trabajador en los años que se indica en cierta comunidad.

Años	IPC
2008	115,1
2009	119,2
2010	121,6
2011	123,8

Años	Incremento salarial anual (%) Anual IPC
2008	
2009	1,8
2010	2,7
2011	1,7

a) Calcula el incremento medio y total del salario en el período 2008-2011.
b) Calcula el incremento anual, medio y total del IPC en el período 2008-2011.
c) Si las condiciones económicas de la comunidad suponemos que no varían, realiza una previsión del valor del IPC para el año 2013.
d) Estudia para cada año y para el período total la pérdida o ganancia del poder adquisitivo y realiza una interpretación de los datos obtenidos.

Solución

a) Calcula el incremento medio y total del salario en el período 2008-2011.

Sería conveniente recordar en este momento el concepto de pérdida o ganancia de poder adquisitivo que puedes encontrar en el ejercicio 8.

Ganancia o pérdida de poder adquisitivo = $\Delta_{poder\ adquisitivo}$ =

$$= \frac{salario\ \ nuevo}{salario\ \ actualizado\ \ según\ \ IPC} = \frac{salario\ \ anterior \cdot (1+\Delta_{salarial})}{salario\ \ anterior \cdot (1+\Delta_{IPC})} = \frac{(1+\Delta_{salarial})}{(1+\Delta_{IPC})}$$

La pérdida o ganancia del poder adquisitivo, pues, se calcula a partir de la comparación de los incrementos anuales del salario ($\Delta_{salario}$) i del IPC (Δ_{IPC},) paralelamente.

A tal fin, como en el enunciado ya disponemos de los incrementos anuales de los salarios, habrá que calcular los incrementos medio y total en el período 2008-2011.

Año	Salarios	Índice
2008		
2009	Ha aumentado 1,8 %	$I_{2008}^{2009} = 1{,}018$
2010	Ha aumentado 2,7 %	$I_{2009}^{2010} = 1{,}027$
2011	Ha aumentado 1,7 %	$I_{2010}^{2011} = 1{,}017$

Para calcular el incremento medio de los salarios del período consideraremos que tenemos 3 índices, y así calcularemos la raíz tercera del producto de estos factores:

$\sqrt[3]{1{,}018 \cdot 1{,}027 \cdot 1{,}017} = \sqrt[3]{1{,}0633} = 1{,}0207$ que interpretamos como que los salarios han aumentado un promedio de 2,07 % anual.

Para calcular el incremento total del período lo haremos a partir de los índices encadenados:

$I_{2008}^{2011} = I_{2010}^{2011} \cdot I_{2009}^{2010} \cdot I_{2008}^{2009} = 1{,}017 \cdot 1{,}027 \cdot 1{,}018 = 1{,}0633$ que interpretamos como que los salarios han aumentado un 6,33 % a lo largo de los tres años.

b) Calcula el incremento anual, medio y total del IPC en el período 2008-2011.

Para calcular el incremento anual, medio y total del IPC en el período 2008-2011, calcularemos la secuencia de índices encadenados que presentamos a continuación, donde podemos ver los incrementos anuales:

Año	IPC	Índice	
2008	115,1		
2009	119,2	$I_{2008}^{2009} = \frac{119{,}2}{115{,}1} = 1{,}0356$	Ha aumentado 3,56 %
2010	121,6	$I_{2009}^{2010} = \frac{121{,}6}{119{,}2} = 1{,}0201$	Ha aumentado 2,01 %
2011	123,8	$I_{2010}^{2011} = \frac{123{,}8}{121{,}6} = 1{,}0181$	Ha aumentado 1,81 %

Para calcular el incremento total del período podemos operar a partir de los índices encadenados:

$$I_{2008}^{2011} = I_{2010}^{2011} \cdot I_{2009}^{2010} \cdot I_{2008}^{2009} = 1{,}0181 \cdot 1{,}0201 \cdot 1{,}0356 = 1{,}0756$$

o también a partir de los valores del IPC iniciales y finales del período total:

$$I_{2008}^{2011} = \frac{123,8}{115,1} = 1,0756$$

interpretaremos que el IPC ha aumentado un 7,56 % en todo el período.

Para calcular el incremento medio del IPC, también podemos operar paralelamente:

$$\sqrt[3]{I_{2010}^{2011} \cdot I_{2009}^{2010} \cdot I_{2008}^{2009}} = \sqrt[3]{1,0181 \cdot 1,0201 \cdot 1,0356} = \sqrt[3]{1,0756} = 1,0246$$

u obrar así:

$$\sqrt[3]{I_{2008}^{2011}} = \sqrt[3]{\frac{123,8}{115,1}} = \sqrt[3]{1,0756} = 1,0246$$

interpretaremos que el IPC ha aumentado un promedio de 2,46 % cada año.

c) Si las condiciones económicas de la comunidad suponemos que no varían, realiza una previsión del valor del IPC para el año 2013.

Partiremos del último dato conocido del IPC del año 2011 y lo incrementaremos con el porcentaje que hemos obtenido como incremento medio en el apartado anterior:

$$IPC_{2013} = IPC_{2011} \cdot (Inc.\ medio\ anual)^2 = 123,8 \cdot 1,0246^2 = 129,97$$

d) Estudia para cada año y para el período total la pérdida o ganancia del poder adquisitivo y realiza una interpretación de los datos obtenidos.

Para calcular las variaciones del poder adquisitivo presentaremos los datos de los incrementos en la siguiente tabla:

Año	2009	2010	2011	Total
Inc. salarial	1,8	2,7	1,7	6,33
Inc. IPC	3,56	2,01	1,81	7,56
Inc. poder adquisitivo	–1,7	+0,68	–0,11	-1,14

Ya hemos explicado al comenzar el ejercicio el concepto de *pérdida o ganancia de poder adquisitivo*. Se gana poder adquisitivo cuando el incremento salarial está por encima del incremento del IPC que nos indica, asimismo, el incremento de los precios de los bienes de consumo. Del mismo modo, habrá una pérdida de poder adquisitivo cuando el incremento salarial esté por debajo del incremento del IPC.

Para hacer esta comparación partiremos de los índices que hemos calculado en las dos tablas anteriores y, como se hace un estudio en términos relativos, haremos

los cocientes de estas cantidades año por año. Mostramos los resultados en la tabla anteriores y los cálculos de las casillas sombreadas en gris están a continuación:

Año 2009 → $\dfrac{1{,}018}{1{,}0356} = 0{,}9830$ → –1,7 %

Año 2010 → $\dfrac{1{,}027}{1{,}0201} = 1{,}0068$ → +0,68 %

Año 2011 → $\dfrac{1{,}017}{1{,}0181} = 0{,}9989$ → –0,11 %

Para estudiar el período total → $\dfrac{1{,}0633}{1{,}0756} = 0{,}9886$ → –1,14 %

Queda más claro si anotamos las interpretaciones en porcentajes, y así lo mostramos en la anterior tabla. Convendremos que el signo positivo indica ganancia de poder adquisitivo y el signo negativo, pérdida.

Ejercicio 10

Para hacer un estudio de la evolución del precio de cierto modelo de ordenador en términos reales, disponemos de los datos que presentamos en la tabla siguiente:

a) Calcula el incremento anual, medio y total del precio del ordenador en términos reales.
b) Si seguimos esta evolución, estima el precio que podría tener el ordenador en 2008.

c)

	2000	2001	2002	2003	2004
IPC base 1992	131	135	139		
IPC base 2002			103	106	109
	1300	1275	1250	1100	950

Nota: Debemos recurrir a períodos y valores muy antiguos o imaginados para trabajar el objetivo del cambio de base del IPC, debido a que con la nueva metodología del cálculo del IPC por el INE esta circunstancia se ha superado, pero es importante que el alumno conozca este contenido para advertir la necesidad de no trabajar en series de IPC no adecuadas en un mismo ejercicio.

Solución

Para empezar, hay que hacer un cambio de base y encontrar todos los valores del IPC en la misma base. Llenaremos la tabla con los nuevos valores encontrados (en rojo) y bajo anotaremos los cálculos realizados.

	2000	2001	2002	2003	2004
IPC base 1992	131	135	139	143,05	147,10
IPC base 2002	97,07	100,04	103	106	109

Veamos los datos del año 2003:

$$IPC_{2002}^{2002} = 103 \rightarrow IPC_{2002}^{2003} = 106$$
$$IPC_{1992}^{2002} = 139 \rightarrow IPC_{1992}^{2003} = X$$

Para calcular el dato desconocido:

$$IPC_{1992}^{2003} = \frac{IPC_{2002}^{2003} \cdot IPC_{1992}^{2002}}{IPC_{2002}^{2002}} = \frac{109 \cdot 139}{103} = 106\frac{139}{103} = 106 \cdot 1{,}35 = 143{,}05$$

Del mismo modo, calculemos el dato del 2004:

$$IPC_{2002}^{2002} = 103 \rightarrow IPC_{2002}^{2004} = 109$$
$$IPC_{1992}^{2002} = 139 \rightarrow IPC_{1992}^{2004} = X$$

Para calcular el dato desconocido:

$$IPC_{1992}^{2004} = \frac{IPC_{2002}^{2004} \cdot IPC_{1992}^{2002}}{IPC_{2002}^{2002}} = \frac{109 \cdot 139}{103} = 106\frac{139}{103} = 109 \cdot 1{,}35 = 147{,}10$$

Podríamos plantear de la misma manera (por proporcionalidad) los cálculos para obtener el resto de datos de la base 2002.

Para continuar el ejercicio utilizaremos los IPC en esta base 1992 o en 2002. Es indiferente siempre y cuando tengamos cuidado de trabajar todos los índices en la misma base.

a) Calcula el incremento anual, medio y total del precio del ordenador en términos reales.

Dado que el análisis del precio del ordenador nos lo piden en términos reales, primero es necesario hacer la conversión del precio del ordenador en moneda constante del año 2000 (deflactación de la serie de precios del ordenador).

Año	Precio ordenador (términos nominales)	Precio ordenador (términos reales 2000)
2000	1300	1300
2001	1275	$1275 \cdot \frac{IPC_{1992}^{2000}}{IPC_{1992}^{2001}} = 1275 \cdot \frac{131}{135} = 1237,2$
2002	1250	$1250 \cdot \frac{IPC_{1992}^{2000}}{IPC_{1992}^{2002}} = 1250 \cdot \frac{131}{139} = 1178,06$
2003	1100	$1100 \cdot \frac{IPC_{1992}^{2000}}{IPC_{1992}^{2003}} = 1100 \cdot \frac{131}{143,05} = 1007,34$
2004	950	$950 \cdot \frac{IPC_{1992}^{2000}}{IPC_{1992}^{2004}} = 950 \cdot \frac{131}{147,10} = 846,02$

Para calcular el incremento anual del precio del ordenador en términos reales, operamos en la tabla siguiente mediante índices encadenados:

Año	Precio ordenador (términos nominales)	Precio ordenador (términos reales 2000)	Índice	Interpretación: Incremento anual
2000	1300	1300	-----	
2001	1275	1237,2	$I_{2000}^{2001} = \frac{1237,2}{1300} = 0,9517$	Ha disminuido un 4,83 %
2002	1250	1178,06	$I_{2001}^{2002} = \frac{1178,06}{1237,2} = 0,9522$	Ha disminuido un 4,78 %
2003	1100	1007,34	$I_{2002}^{2003} = \frac{1007,34}{1178,06} = 0,8551$	Ha disminuido un 14,49 %
2004	950	846,02	$I_{2003}^{2004} = \frac{846,02}{1007,34} = 0,8399$	Ha disminuido un 16,01 %

Para calcular el incremento total y medio del período, operamos:

$I_{2000}^{2004} = \frac{846,02}{1300} = 0,6508$ → A lo largo del período ha disminuido un 34,92 % su valor en términos reales, que equivale a un incremento medio anual de:

$\sqrt[4]{\frac{846,02}{1300}} = \sqrt[4]{0,6508} = 0,8982$ → una disminución anual media del 10,18 %

b) Si seguimos esta evolución, estima el precio que podría tener el ordenador en 2008.

Para hacer este apartado, consideramos que nos piden la estimación en términos corrientes del precio, suponiendo que no varía el comportamiento del IPC ni la evolución del precio del ordenador en términos reales que hemos analizado en el apartado anterior.

Para hacer estas estimaciones necesitamos el incremento medio de las dos series. Nos falta calcular el incremento medio del IPC en el período que nos ocupa:

El IPC ha aumentado un 2,94 % anualmente, por lo que podemos estimar el valor del IPC en el año 2008.

$$IPC_{1992}^{2008} = IPC_{1992}^{2004} \cdot 1{,}0294^4 = 147{,}10 \cdot 1{,}0294^4 = 165{,}18$$

Haremos las mismas operaciones con la serie de los precios de los ordenadores en términos reales.

Precio ordenador para el año 2008 en t. reales del 2000 = $846{,}02 \cdot 0{,}8982^4$ = 550,65 euros.

Para pasarlos a términos corrientes en moneda del 2008:

$$550{,}65 \cdot \frac{IPC_{1992}^{2008}}{IPC_{1992}^{2000}} = 550{,}65 \cdot \frac{165{,}18}{131} = 694{,}3 \text{ €}$$

UNIDAD 4

Series temporales

Introducción teórica

Como elementos introductorios de este capítulo, es conveniente recordar definiciones de conceptos que necesitaremos para alcanzar los objetivos de esta unidad (referencias bibliográficas 1, 21 y 24).

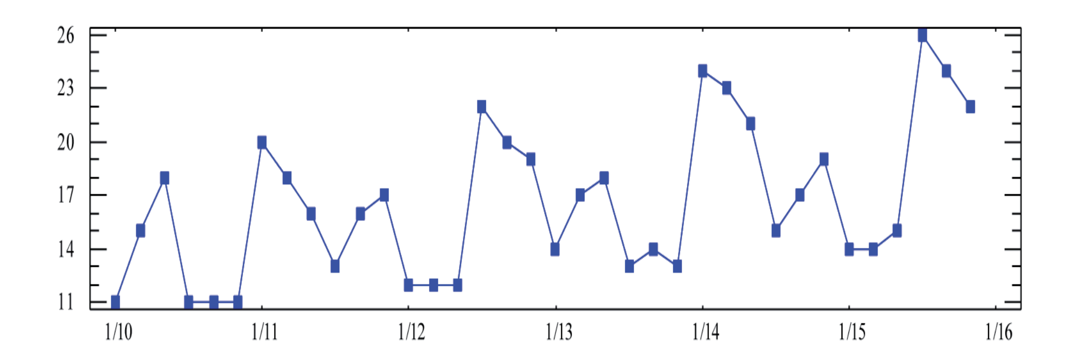
Serie temporal

Es una sucesión de observaciones cuantitativas de un fenómeno ordenadas en los tiempos y períodos equidistantes. Cada observación que denotaremos por y_{ij}, corresponde al valor de la magnitud en el año *i* y período *j*. Ejemplo: si tenemos una serie de observaciones trimestrales a lo largo de los años 2001 a 2005, entenderemos que el dato concierne al valor x_{23} correspondiente al tercer trimestre de 2002.

Gráfica de una serie temporal

La primera herramienta descriptiva que nos puede permitir analizar una serie es su gráfica, que dibujaremos situando los valores de la serie en el eje de ordenadas y los valores de los períodos en el eje de abscisas.

Más adelante detallaremos la importancia de observar este gráfico para asegurarnos de ajustar nuestra serie a un modelo aditivo y para confirmar los resultados de las componentes, que han de reflejar cuantitativamente unos valores que confirmen nuestra visión del fenómeno en la gráfica.



Componentes de una serie temporal

En el análisis de una serie temporal en este tema, consideraremos que toda serie empírica que analicemos está formada por cuatro componentes teóricas: *tendencia, variaciones estacionales, variaciones cíclicas y variaciones residuales*.

Tendencia: es la componente que nos explica el comportamiento del fenómeno a «largo plazo». La denotaremos por T_{ik} y nos permitirá explicar si las medias anuales de los valores de la serie aumentan o disminuyen en el período que queremos analizar.

Variaciones cíclicas: son variaciones que se producen con una periodicidad superior al año y frecuentemente se manifiestan como consecuencia de períodos de prosperidad y de depresión en la actividad económica, o en otras magnitudes cualquiera. Las denotamos por c_{ik}. No las obtendremos en este tema, ya que quedan fuera del alcance de las técnicas que desarrollaremos para tablas pequeñas a fin de estudiar el desarrollo y la justificación teóricos y explícitos de los cálculos. Para tablas más grandes nos ayudaremos de software que nos dará resultados que habrá que analizar con las consideraciones teóricas que podremos ver en los ejercicios propuestos con tablas de menor tamaño.

Variaciones estacionales: son oscilaciones que se producen con una periodicidad dentro del año y que se pueden identificar repetidamente a lo largo de los años de los que disponemos datos por analizar. Por ejemplo, históricamente las series del paro aumentan en invierno y disminuyen en verano, el volumen de ventas de una superficie comercial tiene subidas significativas en períodos de rebajas, etc. Las podremos medir en valores absolutos (componente estacional e_k) o en valores relativos (índices estacionales I_k) respecto a la media global (M = media aritmética de las medias corregidas).

Variaciones residuales o erráticas: ya que los datos son empíricos, es de esperar que de manera natural haya en ellas pequeñas variaciones aleatorias respecto al modelo teórico que pretende analizar la serie con la información del resto de los componentes. Las denotaremos por r_{ik} y también se denominaran residuos. Es necesario que no presentan periodicidad manifiesta y sean de valor reducido. Cuando cualquiera de sus valores nos llame la atención por su valor absoluto respecto al resto, nos indicará un dato que por cualquier motivo no se ajusta al modelo que pretendemos obtener. Hay que analizar su origen: error, efecto producido por una huelga, un accidente, una perturbación meteorológica que habrá que encontrar con la información pertinente al alcance del contexto de la serie estudiada y que intentaremos explicar para justificar la variación de estos datos en particular que se desajustan del modelo.

Nosotros estudiaremos solo series temporales que supondremos que se ajustan al modelo aditivo, circunstancia que se puede comprobar acudiendo a la bibliografía que referimos y que no hemos desarrollado en esta colección de problemas. Por ello podemos considerar que un dato en particular es el resultado de la suma de sus componentes. Así:

$$y_{ij} = T_{ik} + c_{ik} + e_{ik} + r_{ik}$$

Para hacer el análisis de una serie estudiaremos dos métodos: ajuste analítico y el método de las medias móviles. A continuación pondremos el formulario y la notación de cada uno de los métodos que se podrán seguir en los ejercicios resueltos a continuación:

Método del ajuste analítico

Para trabajar este método, es conveniente presentar los cálculos en forma de tabla (véanse ejemplos en los ejercicios resueltos) convenientemente ordenados.

Para calcular la recta de tendencia, en la parte inferior de la tabla de los datos, calcularemos por columnas las medias mensuales de cada año $\overline{y}_i$, los valores de la escala *i*, y en las dos filas inferiores $\overline{y}_i \cdot i$, i^2, que nos permitirá calcular en la columna *totales* la suma de los valores de cada fila.

Hay que explicar que la fila *i* es una escala que crearemos para facilitar la resolución del sistema lineal de dos ecuaciones con dos incógnitas. El procedimiento nos dará resultados óptimos situando el valor del año 0 en la columna central de la tabla en caso de tener un número impar de años en la tabla, o en cualquiera de las columnas adyacentes en caso de un número par de años por estudiar:

Ejemplo:

						TOTALES
Año	2008	2009	2010	2011	2012	
i	-2	-1	0	1	2	0

							TOTALES
Año	2008	2009	2010	2011	2012	2013	
i	-2	-1	0	1	2	3	3

Con estos datos que hemos acumulado en la columna de *totales*, plantearemos y resolveremos el siguiente sistema:

$$\begin{cases} \sum_i \overline{y}_i = Na + b\sum_i i \\ \sum_i \overline{y}_i \cdot i = a\sum_i i + b\sum_i i^2 \end{cases}$$

donde los coeficientes a y b son los coeficientes de la recta de regresión que denominamos recta de tendencia, la formula es $T_i = a + b$, en la que i hace referencia al año que se indica en la escala de las tablas superiores de los ejemplos y *N* es el número de años o columnas que tiene la tabla de los datos.

Esta recta que encontramos no es sino la recta de ajuste lineal por mínimos cuadrados a las medias anuales $\overline{y}_i$.

Por su interpretación nos fijaremos en el signo de su pendiente (coeficiente b) que nos determinará un fenómeno creciente o decreciente, según el signo de b sea

positivo o negativo respectivamente, y el valor del incremento medio anual de la media anual de los valores de la serie.

El valor de la tendencia lo denotaremos por T_i y lo consideraremos constante para todos los datos del año *i*.

Para calcular la componente estacional, trabajaremos las columnas que se pueden ver a la derecha de la tabla original.

En la primera columna podemos encontrar las medias aritméticas de los valores originales de los datos de cada período o fila: $\overline{y}_k = \frac{\sum_i y_{ij}}{N}$.

La columna siguiente corresponde a las medias corregidas, $\overline{y}_k^{'}$, la fórmula de las cuales es:

$$\overline{y}_k^{'} = \overline{y}_k - \frac{b}{m}(k-1)$$

donde *b/m* podemos interpretarlo como el incremento que correspondería a cada período del incremento anual de los datos, debido a la tendencia del fenómeno. Por eso se corrige este incremento con la fórmula antes indicada.

Notamos que *m* es el número de filas de la tabla original, que corresponde al número de observaciones que disponemos en cada año. Así, *m* = 12 si se tratan de observaciones mensuales, *m* = 4 si se trata de observaciones trimestrales, etc.

M es la media global corregida y es la media de las medias corregidas antes definidas:

$$M = \frac{\sum_k \overline{y}_k^{'}}{m}$$

que podemos interpretar como el valor medio de las nuevas medias corregidas y que representará el 100 % o valor de referencia, frente a la que se comparan los comportamientos estacionales que calculamos en las dos columnas de la derecha de la tabla y que son la componente estacional $e_k = \overline{y}_k^{'} - M$ y los índices estacionarios $I_k = \frac{\overline{y}_k^{'}}{M} \cdot 100$. Con estos resultados podremos interpretar en qué períodos los valores de las observaciones están por encima o por debajo del valor de *M*. e_k nos presenta esta desviación en cantidades absolutas, mientras que I_k lo indica de manera porcentual.

Para calcular la componente residual r_{ik} habrá que determinar primeramente el valor de la tendencia T_i para cada año considerado en la tabla, sustituyendo en la recta de tendencia el valor de i correspondiente y consideraremos para la componente estacional e_k los valores que ya hemos calculado y explicado en los párrafos anteriores.

Así, para cada observación, operaremos $r_{ik} = y_{ik} - T_i - e_k$ y dispondremos los resultados en la distribución de la tabla original para facilitar su interpretación e identificación del período y año correspondientes.

Hacemos esta advertencia porque todos sabemos que las cantidades que hay que obtener de la componente residual deberían ser pequeñas en valor absoluto, y que no presentan ninguna regularidad. Ya sabemos que estamos calculando las cantidades no explicadas por nuestro modelo y que permitirán resaltar aquellos valores puntuales que, por razones no predecibles, muestran divergencia del valor que cabría esperar, atendiendo a las componentes de la tendencia y estacional.

Hacer predicciones para los años próximos implica que el análisis de nuestro modelo sea vigente y que ninguna otra circunstancia ajena altere las regularidades que hemos reseñado con nuestro modelo (la tendencia explicada y el comportamiento de los períodos ya cuantificado).

Si queremos prever las cantidades de los próximos años será necesario calcular los valores de su tendencia sustituyendo en la ecuación de su recta los valores de i que les correspondería en caso de que la tabla continuara.

También consideraremos los valores obtenidos de la componente estacional antes mencionada, y podremos hacer las previsiones de los datos futuros operando $y_{ik} = T_i + e_k$.

Método de las medias móviles

Este método está basado en el «suavizado» de una serie cuando esta es sustituida por una sucesión de medias aritméticas de p observaciones, como explicaremos a continuación. En este apartado teórico (cálculo de las medias móviles) cambiaremos la notación de la serie de observaciones inicial y pasaremos a ordenarla con un único subíndice, considerándola como una sucesión ordenada sin contemplar su procedencia de período y año.

Para aplicar este método hay que elegir un número p de observaciones por promediar con unos criterios que después explicaremos.

Si p es impar, formaremos una serie nueva de medias que será:

$$\overline{y}_{\frac{p+1}{2}} = \frac{y_1 + y_2 + \ldots + y_p}{p}, \; \overline{y}_{\frac{p+3}{2}} = \frac{y_2 + y_3 \ldots + y_p + y_{p+1}}{p}, \; \overline{y}_{\frac{p+5}{2}} = \frac{y_3 + y_4 \ldots + y_{p+1} + y_{p+2}}{p}$$

donde puede verse que los subíndices que adjudicamos a estas medias obtenidas $\frac{p+1}{2}$, $\frac{p+3}{2}$, $\frac{p+5}{2}$... corresponden a un número entero, por lo que estas medias podemos hacerlas corresponder a un período original, ya que corresponde al centro de los períodos promediados.

Si p es par, esta circunstancia no se da, ya que $\frac{p+1}{2}$, $\frac{p+3}{2}$, $\frac{p+5}{2}$... no corresponden en este caso a un número entero, por lo que, ante la imposibilidad de hacer corresponder las medias aritméticas en algún período de la serie original, haremos los cálculos de la misma manera, y posteriormente haremos un «centrado» calculando la media aritmética de cada dos medias móviles consecutivas antes calculadas.

Así:

$$\overline{\overline{y}}_{\frac{p+2}{2}} = \frac{\overline{y}_{\frac{p+1}{2}} + \overline{y}_{\frac{p+3}{2}}}{2},\ \overline{\overline{y}}_{\frac{p+4}{2}} = \frac{\overline{y}_{\frac{p+3}{2}} + \overline{y}_{\frac{p+5}{2}}}{2},\ \overline{\overline{y}}_{\frac{p+6}{2}} = \frac{\overline{y}_{\frac{p+5}{2}} + \overline{y}_{\frac{p+7}{2}}}{2} \ldots$$ ya que ahora los nuevos subíndices $\frac{p+2}{2}$, $\frac{p+4}{2}$, $\frac{p+6}{2}$... sí se corresponden a números enteros y, consecuentemente, a períodos concretos de la serie inicial de observaciones.

Este método está basado en que la nueva serie de medias móviles nos permitirá vislumbrar la tendencia de la serie original a «largo plazo», ya que se suaviza el valor individual de cada uno de los datos y las oscilaciones. Para que esta afirmación sea cierta hay que elegir convenientemente el número p de observaciones por promediar, como hemos indicado antes y vamos a detallar a continuación.

El número p es necesario que sea múltiplo del número de observaciones anuales (m) a fin de considerar en cada media todas las fluctuaciones estacionales. Así, debido al método de construcción de las medias móviles antes mencionado, en cada media se sustituirá un dato que corresponde a un cierto período por otra que se corresponde al mismo período en la media siguiente, y siempre tenemos asegurada en cada cálculo la media de todas las oscilaciones de los diferentes períodos dentro de un año o más.

El otro criterio a tener en cuenta se basa en la observación de la gráfica de la serie original, la importancia de la cual ya hemos comentado al comenzar este apartado teórico. Observando esta gráfica hay que intentar encontrar una cierta periodicidad superior al año, es decir, hay que anular el efecto de una componente estacional, tomando un número p que ha de ser múltiplo del número de observaciones q que comprende el «período» gráfico que se repite a lo largo de la serie.

De ello resulta que, para calcular la tendencia, habrá que coger un número p de observaciones que sea el mínimo común múltiplo de m y q. Por ejemplo, si tenemos una serie de datos trimestrales, m = 4 y de la observación de la gráfica podemos ver un patrón bianual que se repite aproximadamente cada 8 observaciones, en ese caso el número p que deberemos considerar para el cálculo de las medias móviles

será 8, y por tratarse de un número par será necesario hacer después un posterior «centrado».

Para calcular la componente estacional, determinaremos por un lado la media aritmética de todos los datos que corresponden a cada período y que denotaremos por $\overline{y_k}$, por otro, calculamos las medias móviles con $p = m$, número de observaciones anuales. En caso de que la serie tenga una componente cíclica anual, podremos aprovechar los cálculos del apartado de la tendencia.

A continuación dispondremos las medias móviles en la distribución bidimensional que originariamente tenían los datos, distribuyéndolas en años por columnas y en períodos por filas. Podremos observar que el método de las medias móviles centradas obliga a que algunas de las celdas de la tabla queden vacías, ya que no podemos hacer corresponder ningún dato a los períodos iniciales y finales.

A continuación, calcularemos la media aritmética de las medias que corresponden a cada período y que denotaremos por E_k, ya que podemos considerarla como la componente extraestacional, pues el comportamiento estacional ha sido anulado por la elección de los datos conveniente para hacer las medias móviles con $p = m$.

Así: $\overline{e_k} = \overline{y_k} - \overline{E_k}$.

Para finalizar esta parte, recordamos que los temas siguientes serán el desarrollo de la probabilidad y la inferencia (referencias bibliográficas 7 y 26).

Objetivos

Los problemas deben permitir que los alumnos alcancen los objetivos didácticos:

4*a*) Reconocer en una colección de datos los patrones y la notación de una serie temporal.
4*b*) Analizar una serie y a partir de su gráfica, poder comprobar que se adapta al modelo aditivo.
4*c*) Conocer las diferentes componentes de una serie temporal: tendencia, componente estacional, componente errática o residual.
4*d*) Saber calcular las diferentes componentes de una serie, suponiendo un modelo aditivo, por el método del ajuste analítico.
4*e*) Saber calcular las diferentes componentes de una serie, suponiendo un modelo aditivo, por el método de las medias móviles.
4*f*) Saber interpretar los resultados obtenidos de las diferentes componentes de una serie temporal y relacionarlos con la gráfica de la serie.
4*g*) Hacer estimaciones de los valores de una serie temporal en fechas futuras cercanas a los valores analizados, en el ajuste analítico.

Objetivos *Ejercicios*	*4a*	*4b*	*4c*	*4d*	*4e*	*4f*	*4g*
1	x	x	x	x		x	x
2	x	x	x	x		x	x
3	x	x	x	x		x	x
4	x	x	x				
5	x	x	x		x	x	
6	x	x	x		x	x	
7	x	x	x	x	x	x	x

Enunciados

4*a*) Reconocer en una colección de datos los patrones y la notación de una serie temporal.
4*b*) Analizar una serie y a partir de su gráfica, poder comprobar que se adapta al modelo aditivo.
4*c*) Conocer las diferentes componentes de una serie temporal: tendencia, componente estacional, componente errática o residual.

Ejercicio 1

Para analizar la evolución de los gastos en un departamento de una empresa, se tomaron los siguientes datos, que expresan en miles de euros los gastos cuatrimestrales de los cuatro años que figuran en la tabla:

	2008	2009	2010	2011
1.er cuatrimestre	26	25	21	20
2.o cuatrimestre	18	15	12	10
3.er cuatrimestre	22	20	18	12

a) Suponiendo modelo aditivo, calcula por el método del ajuste analítico, las componentes de esta serie e interpreta cada uno de los resultados obtenidos.
b) Estima los valores de los gastos que se pueden esperar para el año 2012.

4*a*) Reconocer en una colección de datos los patrones y la notación de una serie temporal.
4*b*) Analizar una serie y a partir de su gráfica, poder comprobar que se adapta al modelo aditivo.
4*c*) Conocer las diferentes componentes de una serie temporal: tendencia, componente estacional, componente errática o residual.

Ejercicio 2

Los siguientes datos, extraídos del INI, nos muestran las pernoctaciones hoteleras en la Comunidad Valenciana, desde el año 2006 hasta el 2010.

Realiza un análisis del fenómeno, obteniendo las diferentes componentes de la serie temporal para el ajuste analítico, suponiendo un modelo aditivo. Interpreta el significado de cada una de las componentes.

Haz las previsiones que podemos esperar para los años 2012 y 2013.

	2006	2007	2008	2009	2010
Enero	1.296.648	1.307.384	1.289.627	1.115.865	1.061.450
Febrero	1.424.453	1.472.806	1.562.471	1.329.013	1.347.885
Marzo	1.750.282	1.892.925	1.993.175	1.700.733	1.800.323
Abril	2.152.783	2.226.734	1.868.049	1.925.454	1.943.080
Mayo	2.131.194	2.181.889	2.114.631	1.952.409	2.049.616
Junio	2.399.782	2.523.555	2.311.444	2.241.147	2.049.616
Julio	2.884.491	2.983.227	2.877.028	2.871.011	2.945.899
Agosto	3.153.407	3.308.833	3.227.055	3.274.561	3.341.961
Septiembre	2.540.711	2.580.090	2.524.888	2.381.776	2.424.567
Octubre	2.207.203	2.097.005	2.012.838	1.954.615	2.053.500
Noviembre	1.674.433	1.741.296	1.498.967	1.471.579	1.472.610
Diciembre	1.437.036	1.420.988	1.251.809	1.201.342	1.174.451

4*a*) Reconocer en una colección de datos los patrones y la notación de una serie temporal.
4*b*) Analizar una serie y a partir de su gráfica, poder comprobar que se adapta al modelo aditivo.
4*c*) Conocer las diferentes componentes de una serie temporal: tendencia, componente estacional, componente errática o residual.

Ejercicio 3

Con los siguientes datos, extraídos de la DGT, que nos muestran las nuevas licencias de todos los tipo de carnés de conducir en la Comunidad Valenciana, desde el año 2008 hasta el 2010, realiza un análisis del fenómeno, obteniendo las diferentes componentes de la serie temporal para el ajuste analítico.

Interpreta el significado de cada una de las componentes.

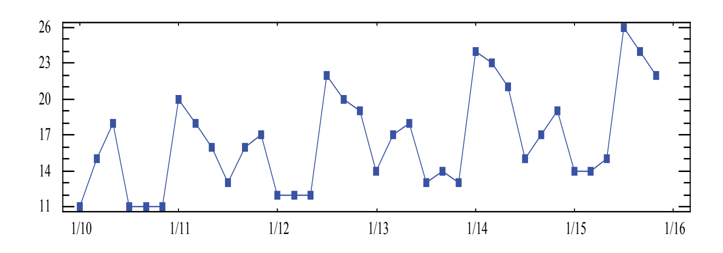
Haz las previsiones que podemos esperar para los años 2011 y 2012.

	2008	2009	2010
Enero	12031	8380	7071
Febrero	12208	8993	7685
Marzo	9497	7973	8444
Abril	12862	7360	6781
Mayo	12567	7874	7728
Junio	12723	7881	7585
Julio	19003	11820	11138
Agosto	2147	1346	2205
Septiembre	10826	8876	7901
Octubre	11196	8374	7427
Noviembre	10628	10137	7665
Diciembre	9064	8083	5746

4*a*) Reconocer en una colección de datos los patrones y la notación de una serie temporal.
4*b*) Analizar una serie y a partir de su gráfica, poder comprobar que se adapta al modelo aditivo.
4*c*) Conocer las diferentes componentes de una serie temporal: tendencia, componente estacional, componente errática o residual.

Ejercicio 4

La siguiente gráfica es la representación de una serie temporal donde se detallan los datos bimensuales de 6 años. Si tuviéramos que calcular la tendencia y la componente estacional de dicha serie por el método de las medias móviles, explica la elección del número de datos que hay que considerar para el cálculo de las medias (*p*) en cada caso, justificando la respuesta.



4*a*) Reconocer en una colección de datos los patrones y la notación de una serie temporal.
4*b*) Analizar una serie y a partir de su gráfica, poder comprobar que se adapta al modelo aditivo.

Ejercicio 5

En la siguiente tabla presentamos el número total de viajeros transportados en los servicios de transporte público en la Comunidad Valenciana, detallados por meses, de los años 2006 al 2010.

	2006	2007	2008	2009	2010
Enero	12340	12542	12622	11869	10283
Febrero	11850	12156	12202	11365	10981
Marzo	13721	13729	10457	12024	12207
Abril	10919	11612	12960	10469	10549
Mayo	13495	13777	12713	12038	12076
Junio	13029	13094	12619	12314	11783
Julio	12118	12364	12351	11490	10621
Agosto	8803	8814	8846	8027	7698
Septiembre	12148	11768	12057	10964	10705
Octubre	13141	13266	13277	11956	11447
Noviembre	13307	12655	12474	11718	11557
Diciembre	11859	11624	11682	10894	10769

a) Calcula las componentes de esta serie, por el método de las medias móviles.

Interpreta los resultados.

P. Juan Verdoy / M. J. Beltrán / M. J. Peris - ISBN: 978-84-15444-38-1 Problemas resueltos de estadística aplicada a las ciencias sociales - UJI - DOI: http://dx.doi.org/10.6035/Sapientia100

4*a*) Reconocer en una colección de datos los patrones y la notación de una serie temporal.
4*b*) Analizar una serie y a partir de su gráfica, poder comprobar que se adapta al modelo aditivo.

Ejercicio 6

Realiza la gráfica de la siguiente serie que indica los miles de kilos de fruta comercializada por trimestres en los últimos 4 años.

	2008	2009	2010	2011
1.er trimestre	10	23	12	29
2.o trimestre	11	27	11	28
3.er trimestre	9	25	10	21
4.o trimestre	8	20	8	23

Observa las siguientes tablas que presenten los cálculos que hemos hecho para obtener la tendencia y la componente estacional de dicha serie.

Identifica el método empleado, añade aquellos datos que faltan en las tablas, comentando el procedimiento de cálculo que hay que hacer en las mesas, justificándolos.

datos		
10		
11		
9		
8		
23		
27		
	16,875	
25		16,9375
	17	
20		17
	17	
12		17,375
11		
	17,875	
10		17,625
	17,375	
8		
29		
28		
21		
23		

datos		
10		
11		
9		
	12,75	
8		
	16,75	
23		18,75
27		
	23,75	
25		22,375
	21	
20		19
	17	
12		15,125
	13,25	
11		11,75
	10,25	
10		
8		16,625
	18,75	
29		20,125
28		
21		
23		

4*a*) Reconocer en una colección de datos los patrones y la notación de una serie temporal.
4*b*) Analizar una serie y a partir de su gráfica, poder comprobar que se adapta al modelo aditivo.
4*c*) Conocer las diferentes componentes de una serie temporal: tendencia, componente estacional, componente errática o residual.
4*d*) Saber calcular las diferentes componentes de una serie, suponiendo un modelo aditivo, por el método del ajuste analítico.

Ejercicio 7

La siguiente serie cronológica muestra el número de nuevas contrataciones de una superficie comercial por cuatrimestres en los años indicados:

	2007	2008	2009	2010	2011
1.er cuatrimestre	41	39	35	21	22
2.o cuatrimestre	37	33	27	15	16
3.er cuatrimestre	36	30	16	12	13

a) Realiza la gráfica de la serie y, suponiendo un modelo aditivo, calcula por el método del ajuste analítico las componentes de esta serie e interpreta cada uno de los resultados obtenidos.
b) Haz las previsiones que podemos esperar para los años 2012 y 2013.
c) Realiza el análisis de la serie por el método de las medias móviles, estimando también las componentes de la serie.
d) Compara los resultados del análisis por los dos métodos empleados.

Ayudas

En este apartado se presentarán las ayudas para emplear en caso de ser necesario a la hora de realizar los ejercicios y problemas. Es conveniente no hacer un abuso excesivo de estas ayudas, es decir, antes de emplear la ayuda hay que pensar el problema al menos durante unos 10-15 minutos. Después se consultará la ayuda de tipo 1 y se intentará resolver el ejercicio con esta ayuda. Si no es posible resolverlo, entonces se consultará la ayuda de tipo 2, y en último término la solución.

Ayudas Tipo 2

Ejercicio 1

Para resolver este problema sería conveniente llenar la siguiente tabla. En la parte inferior de los datos realizaremos los cálculos referentes a la recta de tendencia. Con los valores de las celdas de los totales podremos resolver el sistema correspondiente.

Observa la diferencia en este apartado por ser un número impar de columnas (en el ejercicio 2 había 4 columnas) y para hacer la escalera tal como está indicada en la siguiente tabla (fila i).

En las columnas de la derecha podremos hacer los cálculos para obtener la componente estacional y los índices estacionales (en porcentaje).

	2006	2007	2008	2009	2010		$\overline{y}_k$	$\overline{y}_k'$	e_k	I_k
						TOTALES				
$\overline{y}_i$								M =		
i	-2	-1	0	1	2					
$\overline{y}_i \cdot i$										
i^2										

Recuerda interpretar los resultados que te permitirán confirmar la coherencia de tus resultados con los datos originales.

Como se trata de una serie mensual con muchos datos habría que ayudarse con una hoja de cálculo.

Para hacer las previsiones habrá que calcular previamente la tendencia de cada año y la componente de cada período, y para obtener los datos que queremos, sumaremos en cada caso los datos correspondientes.

Ejercicio 2

Para hacer este ejercicio seguiremos el mismo procedimiento que en el ejercicio 2. También se trata de una serie mensual, por lo que sería conveniente utilizar una hoja de cálculo. Por tratarse también de un número impar de columnas y utilizando la escala de la fila *i*, poniendo el 0 en la columna central, se simplifica mucho la resolucion del sistema que hay que plantear para la tendencia.

Tal como se ha comentado en el ejercicio 2, es conveniente trabajar los datos rellenando una tabla similar a la del ejercicio anterior.

Ejercicio 3

En este ejercicio no se pide hacer ningún cálculo. Tan solo es una reflexión para aprender a elegir el número de datos (p) que hay que coger para hacer las medias móviles, tanto para el cálculo de la tendencia como para el cálculo de la componente estacional.

Ejercicio 4

Se trata de una serie con muchos datos, por lo que habría que ayudarse de una hoja de cálculo porque debemos abordar el análisis por el método de las medias móviles.

Hay que hacer, en primer lugar, la gráfica de los datos y buscar la periodicidad de dicha gráfica. En este caso tiene un comportamiento que se repite año tras año, habrá pues que calcular medias móviles de 12 datos ($p = 12$) para el cálculo de la tendencia. Por tratarse de un número par tendremos que hacer después un posterior «centrado».

Para calcular la componente estacional podremos aprovechar las mismas medias del apartado anterior por tratarse de $p = 12$.

Para trabajar más claramente sería conveniente preparar una tabla parecida a esta (presentamos tan solo los primeros cálculos y espacios).

Las gráficas pueden servirnos también para la interpretación de los resultados y la validación de la coherencia de estos con los datos originales.

Enero 2006	12.340		
Febrero 2006	11.850		
Marzo 2006	13.721		
Abril 2006	10.919		
Mayo 2006	13.495		
Junio 2006	13.029		
		12227,5	
Julio 2006	12.118		12235,9167
		12244,3333	
Agosto 2006	8.803		12257,0833
		12269,8333	

Ejercicio 5

En este ejercicio ya nos dan las tablas preparadas para el método de las medias móviles y con muchos cálculos ya realizados. Es un ejercicio de consolidación de la técnica y también se trabaja el mismo objetivo del problema 4, que es aprender a averiguar el número de datos que hay que tomar para calcular la tendencia y la componente estacional. Comienza pensando qué es este número p en cada caso y completa las tablas donde faltan valores al principio, en medio y al final. Por tratarse de números pares hay que hacer después un «centrado» en cada caso.

Luego también hay que calcular la componente residual.

Ejercicio 6

En realidad este ejercicio es «doble» ya que nos piden el análisis por los dos métodos. Este es un ejercicio para terminar el tema y consolidar nuestro trabajo repasando las dos técnicas de análisis que hemos visto a lo largo del tema.

Habrá que hacer, en primer lugar, un análisis por el método del ajuste analítico. Puedes seguir la pauta de los ejercicios 2 y 3 en este apartado.

Para hacer el análisis por el método de las medias móviles puedes seguir la pauta de los ejercicios 5 y 6 pero con menos datos, lo que simplificará los cálculos.

Nos piden que comparemos ambos métodos para reflexionar sobre las semejanzas y diferencias entre ellos.

Ayudas Tipo 2

Ejercicio 1

Para realizar el análisis de una serie temporal por el método del ajuste analítico es conveniente disponer los datos en la tabla siguiente, como ya se indicaba en la de la ayuda tipo 1.

	06	07	08	09	10		$\overline{y}_k$	$\overline{y}'_k$	e_k	I_k
						TOTALES				
$\overline{y}_i$						10200467		M = 2.059.542		
i	–2	–1	0	1	2	0				
$\overline{y}_i \times i$						–424346				
i^2	4	1	0	1	4	10				

Para calcular la tendencia, nos fijaremos en las celdas de la parte inferior. Hemos llenado los valores de la columna de los «totales» para poder comprobar sus cálculos y habrá que resolver el siguiente sistema:

$$\begin{cases} \sum_i \bar{y}_i = Na + b\sum_i i \\ \sum_i \bar{y}_i \cdot i = a\sum_i i + b\sum_i i^2 \end{cases} \Rightarrow \begin{cases} 10200467 = 5a + 0b \\ -424346 = 0a + 10b \end{cases}$$

Los resultados de este sistema nos permiten obtener la recta de tendencia:

$$T_i = a + b \ = 2040093{,}4 - 42434{,}6i$$

Para calcular la componente estacional, nos fijaremos en las celdas de la parte de la derecha. Recordemos que hay que calcular *b/m* = –42434,6/12 = –3536,2 para poder obtener las medias corregidas. Indicamos el valor de *M* en la tabla.

La componente estacional la obtenemos en valores absolutos y como índice, donde el valor de *M* es el valor de referencia.

Para obtener la componente residual $r_{ik} = y_k - T_i - e_k$, por lo que hay que tener previamente calculados los valores de la tendencia (para cada año) y el valor de la componente estacional (para cada período).

Para hacer las previsiones hay que calcular la tendencia que corresponderá a los años 2012 y 2013, y construir la serie para los períodos que nos piden sumando estas componentes: $y_{ik} = T_i + e_k$.

Ejercicio 2

Este ejercicio es muy parecido al anterior y la tabla y los cálculos seguirán el mismo planteamiento.

La tabla para llenar con los «totales» será la siguiente:

	08	09	10		$\bar{y}_k$	$\bar{y}'_k$	e_k	I_k
				TOTALS		M = 9.772		
$\bar{y}_i$				26602				
i	–1	0	1	0				
$\bar{y}_i \times i$				–3948				
i^2	1	0	1	2				

Algunos resultados parciales que puedes comprobar:

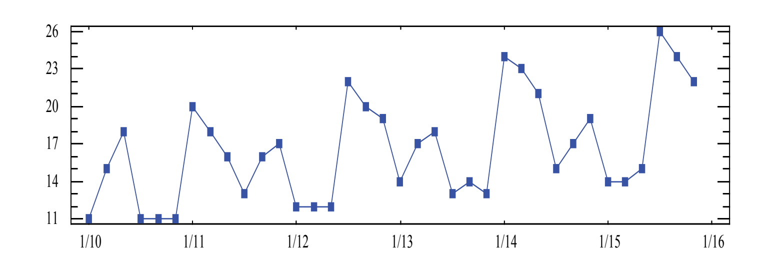
Recta de tendencia: $T_i = a + bi = 8867{,}36 - 1974i$
b/m = –164,5

$e_1 = -611$	$e_2 = 21$	$e_3 = -805$
$e_4 = -278$	$e_5 = 276$	$e_6 = 447$
$e_7 = 5.202$	$e_8 = -6.721$	$e_9 = 745$
$e_{10} = 707$	$e_{11} = 1.350$	$e_{12} = -332$

Con estos valores se debe encontrar la componente errática o residual y las previsiones que nos piden.

Ejercicio 3

Para resolver este ejercicio es necesario observar la gráfica de la serie y el número de datos que hay en cada «período» de la gráfica.



En este caso, para calcular la tendencia debemos considerar que en esta gráfica se «repite» el patrón cada 9 observaciones y que por tratarse de datos bimensuales tenemos 6 datos por año. ¿Cuál será, pues, el número de datos que hay que tomar en cada media para hacer un «suavizado» por el método de las medias móviles?

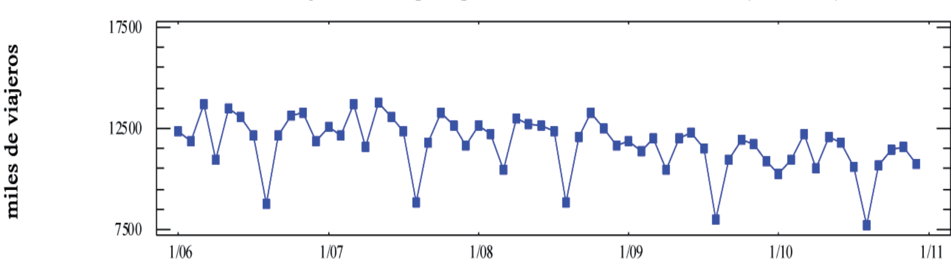
Para calcular la componente estacional, ¿cuál será el número de datos a coger? ¿cuántos datos tenemos por año?

Ejercicio 4

Se trata de una serie con muchos datos por lo que hay que advertir que sería recomendable hacer los cálculos con la ayuda de una hoja de cálculo.

Como vamos a utilizar el método de las medias móviles, en primer lugar hay que decidir el número de datos por considerar en cada media, por lo que hay que ver la gráfica de la serie:





Podemos ver que tiene una periodicidad anual y como tenemos datos mensuales, pues hay que calcular medias con 12 datos. Por tratarse de un número par de datos, tendremos que hacer después un «centrado».

En la siguiente tabla se indican los primeros resultados que podréis comprobar:

Enero 2006	12340		
Febrero 2006	11850		
Marzo 2006	13721		
Abril 2006	10919		
Mayo 2006	13495		
Junio 2006	13029		
		12227,5	
Julio 2006	12118		12235,9167
		12244,3333	
Agosto 2006	8803		12257,0833
		12269,8333	
Septiembre 2006	12148		12270,1667
		12270,5	
Octubre 2006	13141		12299,375
		12328,25	
Noviembre 2006	13307		12340
		12351,75	
Diciembre 2006	11859		12354,4583
		12357,1667	
Enero 2007	12542		
		....	
Febrero 2007	12.156		....
		....	

La columna de la derecha recoge los valores de la tendencia.

Para hallar la componente estacional, calculamos las medias de los datos originales, por períodos y las denotamos como $\bar{y}_k$:

	2006	2007	2008	2009	2010	$\bar{y}_k$
Enero	12340	12542	12622	11869	10283	
Febrero	11850	12156	12202	11365	10981	
Marzo	13721	13729	10457	12024	12207	
Abril	10919	11612	12960	10469	10549	
Mayo	13495	13777	12713	12038	12076	
Junio	13029	13094	12619	12314	11783	
Julio	12118	12364	12351	11490	10621	
Agosto	8803	8814	8846	8027	7698	
Septiembre	12148	11768	12057	10964	10705	
Octubre	13141	13266	13277	11956	11447	
Noviembre	13307	12655	12474	11718	11557	
Diciembre	11859	11624	11682	10894	10769	

Por otra parte, en este caso, podemos aprovechar los resultados de las medias de la tendencia (por estar calculadas con 12 datos, observaciones que tenemos de cada año), aunque hay que redistribuirlos en años y períodos, en forma de tabla y calcular sus medias por períodos (filas). Las denotaremos por E_k.

En la siguiente tabla tienes algunos valores para comprobar, aunque hay otros que deberás calcular.

	2006	2007	2008	2009	2010	E_k
Enero					11041,125	11776,9896
Febrero					10991,2083	11749,8854
Marzo					10966,7083	11723,3437
Abril					10934,7083	11690,6667
Mayo					10906,7917	11654,7917
Junio					10894,875	

Julio	12235,9167					
Agosto	12257,0833					
Septiembre	12270,1667					
Octubre	12299,375					
Noviembre	12340					
Diciembre	12354,4583					

Para calcular la componente estacional habrá que restar las columnas finales de estas dos tablas, $e_k = \bar{y}_k - E_k$, que nos permitirá explicar cuáles son los meses de mayor y menor utilización del transporte público en la Comunidad Valenciana.

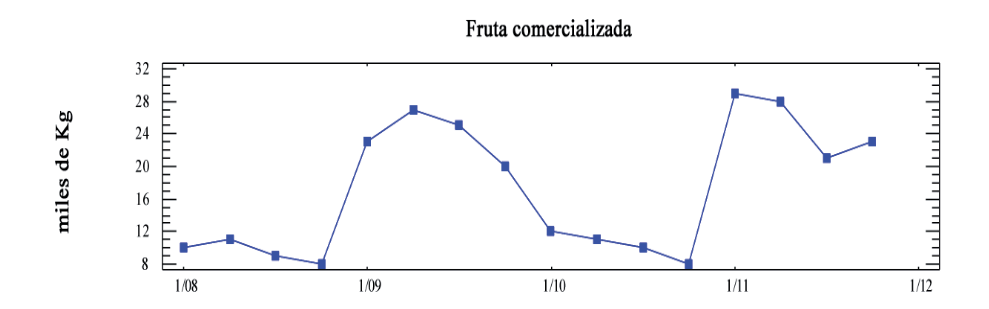
Para calcular la componente errática o residual hay que restarle a cada valor inicial el valor de la tendencia correspondiente a cada celda en los casos en que esta existe (no a los primeros y últimos valores de la serie), y también le restamos el valor de la componente estacional que corresponde a cada período.

En la siguiente tabla se presentan los resultados de estos cálculos que puedes comprobar. ¿Sabrías explicar lo que nos dicen los valores reseñados en rojo y en azul?

	2006	2007	2008	2009	2010
Enero		20,3729	462,9979	20,1646	–912,335
Febrero		–183,0396	235,5021	–220,5396	28,8771
Marzo		661,993725	–2266,21458	–225,214575	536,035425
Abril		–351,25835	1329,90835	–586,50835	3,15835
Mayo		281,82497	–476,38333	–484,84163	4,19997
Junio		–141,799975	–342,841675	77,908325	–54,466675
Julio	20,168725	215,335425	498,793725	433,502125	
Agosto	–77,82652	–101,74322	298,21508	290,75678	
Septiembre	106,939145	–160,477455	331,647545	88,980845	
Octubre	–59,283964	287,674336	460,132736	–59,992264	
Noviembre	302,245684	–99,004316	25,203984	–66,754316	
Diciembre	–224,748979	–130,415679	209,501021	65,251021	

Ejercicio 5

En este ejercicio nos presentan dos tablas que, evidentemente, se corresponden al método de las medias móviles.



Miramos la gráfica de la serie (el patrón de periodicidad se puede decir que se repite cada dos años, 8 observaciones) y como se trata de datos trimestrales (4 datos por año), debemos tomar 8 datos para calcular las medias de la tendencia , ya que mcm (8,4) = 8.

En la tabla de la izquierda faltan valores para llenar al principio, en medio y al final. También faltan por calcular las celdas de la columna de la derecha que corresponde al «centrado» (media aritmética de cada dos valores de la columna anterior).

Mientras, en la tabla de la derecha que nos proponen en el ejercicio, las medias que se han calculado son de 4 observaciones, ya que por tratarse de datos trimestrales tenemos 4 por año. Podemos comprobar que faltan los resultados de algunas celdas en la columna de las medias como la del «centrado» de la columna de la derecha. Estas medias nos permitirán calcular la componente estacional.

Habrá que repetir el proceso del problema anterior:

a) Calcula las medias de los datos originales por períodos $\bar{y}_k$.
b) Organiza en forma de tabla de doble entrada (años y trimestres) los resultados de la última columna de la tabla de la derecha.
c) Calcula las medias para cada período de la tabla anterior, E_k.
d) Los valores de la componente estacional se obtienen restando $e_k = \bar{y}_k - E_i$.
e) Interpreta los resultados.

Para calcular la componente residual o errática: $r_k = y_k - T_k - e_i$. Tan solo podemos determinarla para los dos años centrales por no disponer de más datos de tendencia.

Ejercicio 6

Este ejercicio propone hacer el análisis de la serie por los dos métodos para invitarnos a confrontar las dos técnicas.

a) Por el método del ajuste analítico:

Habrá que rellenar la siguiente tabla. Para el cálculo de la tendencia necesitaremos las celdas de la parte inferior y utilizaremos la columna de «totales» para plantear el sistema que nos permite obtener los coeficientes de la recta de tendencia:

	2007	2008	2009	2010	2011		$\overline{y}_k$	$\overline{y}_k^{'}$	e_k	I_k
1r cuat.	41	39	35	21	22					112,06 %
2n cuat.	37	33	27	15	16					97,87 %
3r cuat.	36	30	16	12	13	TOTALES ↓				90,07 %
$\overline{y}_i$	38	34	26	16	17	131		M = 28,2		
i						0				
$\overline{y}_i \times i$						-60				
i^2						10				

En la tabla presentamos algunos resultados que podremos comprobar, y a continuación los resultados que hay que obtener con los valores de la tabla:

Recta de tendencia: $T_i = 26,-6i$

$b/m = -6/3 = -2$

$$M = \frac{\sum_k \overline{y}_k^{'}}{m} = 28,2$$

Componente estacional:

$$e_1 = \overline{y}_1^{'} - M = 31,6 - 28,2 = 3,4$$
$$e_2 = \overline{y}_2^{'} - M = 27,6 - 28,2 = -0,6$$
$$e_3 = \overline{y}_3^{'} - M = 25,4 - 28,2 = -2,8$$

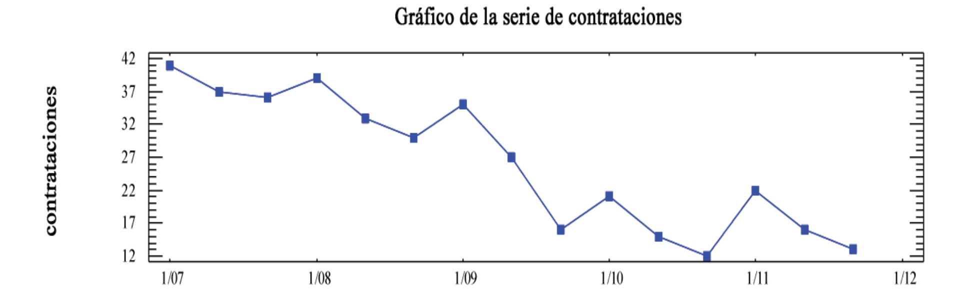
Índices estacionales:

$$I_1 = \frac{\overline{y}_1^{'}}{M} \cdot 100 = \frac{31,6}{28,2} \cdot 100 = 112,05\ \%$$
$$I_2 = \frac{\overline{y}_2^{'}}{M} \cdot 100 = \frac{27,6}{28,2} \cdot 100 = 97,87\ \%$$
$$I_3 = \frac{\overline{y}_3^{'}}{M} \cdot 100 = \frac{25,4}{28,2} \cdot 100 = 90,07\ \%$$

Para calcular la componente residual: $r_{ik} = y_{ik} - T_i - e_k$

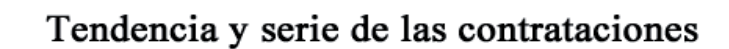
b) Por el método de las medias móviles:

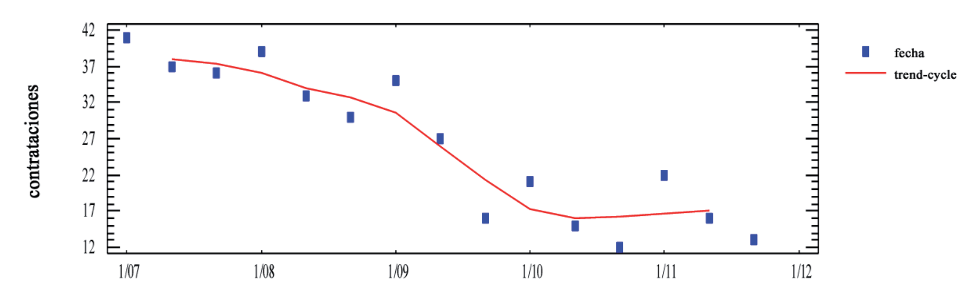
Hay que ver la gráfica de la serie y como su comportamiento es anual y tenemos datos cuatrimestrales, tomaremos 3 datos para calcular cada media y obtener la columna de las medias móviles. Como se trata de un número impar de datos no es necesario hacer un posterior «centrado».



En la tabla siguiente figuran algunos resultados de los valores por comprobar. Faltan otras celdas por llenar. La columna, ya completada, son los valores de la tendencia. Puedes ayudarte de este gráfico para interpretarla:

datos	T_{ik}
41	
37	38,00
36	37,33
39	
33	
30	32,67
35	3067
27	
16	
21	17,33
15	16,00
12	
22	
16	17,00
13	





Estos resultados, que hemos obtenido de las medias, los colocamos de nuevo en forma de tabla de doble entrada (como el enunciado) haciendo corresponder a cada período de la tabla su media, y repetiremos todo el procedimiento como se ha indicado en el ejercicio 6.

La comparación de los dos métodos puedes encontrarla dando respuesta a estas cuestiones, entre otras posibles que tú puedes añadir:

a) ¿Los resultados son iguales? ¿Y su interpretación?
b) ¿Son igualmente fiables o la calidad de las interpretaciones depende de las características de los datos (periodicidad, regularidad, etc.)?
c) ¿Puedes hacer previsiones con ambos métodos?

El ejercicio que debes haber encontrado es el ejercicio 6.

Las portadas se diferencian únicamente por las noticias que aparecen, y no por la posición de cada noticia en la primera plana. Por lo tanto, hay que contar combinaciones de 6 noticias agrupadas de 4 en 4.

Soluciones

Ejercicio 1

Para analizar la evolución de los gastos en un departamento de una empresa, se tomaron los siguientes datos que expresan en miles de euros los gastos cuatrimestrales de los cuatro años que figuran en la tabla:

	2008	2009	2010	2011
1.ᵉʳ cuatrimestre	26	25	21	20
2.º cuatrimestre	18	15	12	10
3.ᵉʳ cuatrimestre	22	20	18	12

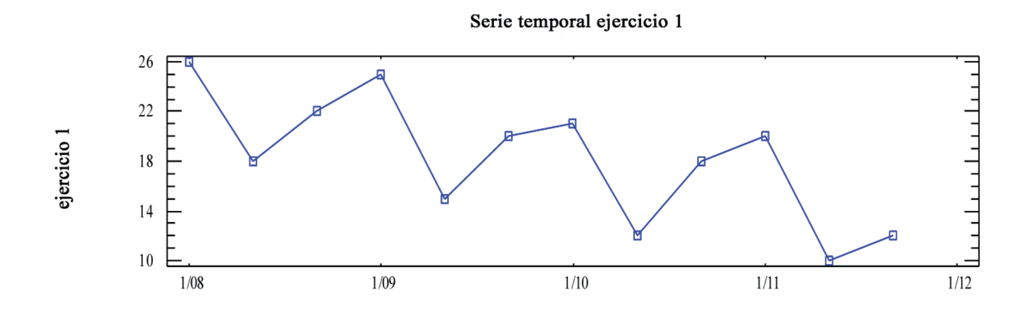
a) Suponiendo un modelo aditivo, calcula por el método del ajuste analítico, las componentes de esta serie e interpreta cada uno de los resultados obtenidos.
b) Estima los valores de los gastos que se pueden esperar para el año 2012.

Solución

a) Suponiendo modelo aditivo, calcula por el método del ajuste analítico, las componentes de esta serie e interpreta cada uno de los resultados obtenidos.

Identificación del modelo y gráfica

Siempre hay que empezar con una gráfica de los datos para ver el patrón de comportamiento de la serie y confirmar que podemos aplicarle un ajuste analítico de tipo aditivo:



Viendo esta gráfica solo cabe esperar una tendencia decreciente y también una componente estacional bastante marcada por la «periodicidad», que podemos ver en la gráfica (eje X) que coincide con el intervalo anual que corresponde cada tres datos.

Cálculo de la tendencia

Para calcular la componente de la tendencia llenaremos las casillas de la parte inferior de la tabla siguiente para ajustar una recta de regresión a las medias cuatrimestrales anuales que figuran en la tabla en la fila $\bar{y}_i = \frac{\sum_k y_{ik}}{m}$.

Hacer notar que $\bar{y}_i$ es la media de los valores de cada columna. Recordemos tan solo que cada dato de una serie temporal la representamos por dos subíndices y_{ik} donde i hace referencia al año y k hace referencia al período, m indica el número de filas de la tabla o número de períodos de los que tenemos datos para cada año. En nuestro caso $m = 3$ porque un año tiene 3 cuatrimestres.

Vemos que en la fila siguiente podemos encontrar una escalera i que hace referencia a los años para simplificar los cálculos; es recomendable poner el valor 0 en alguna de las columnas centrales de la tabla porque simplificará muchos los cálculos posteriores.

A continuación llenaremos las filas tercera y cuarta que hacen referencia a $\bar{y}_i \cdot i$ y i^2, considerando los valores de las filas anteriores.

Después sumaremos las filas y obtendremos los valores que podemos encontrar en la columna «totales».

	2008	2009	2010	2011	
1.er cuatrimestre	26	25	21	20	
2.o cuatrimestre	18	15	12	10	
3.er cuatrimestre	22	20	18	12	TOTALES
$\bar{y}_i$	22	20	17	14	73
i	-1	0	1	2	2
$\bar{y}_i \times i$	-22	0	17	28	23
i^2	1	0	1	4	6

Utilizaremos estos valores «totales» para resolver el sistema que se plantea a continuación:

$$\begin{cases} \sum_i \bar{y}_i = Na + b\sum_i i \\ \sum_i \bar{y}_i \cdot i = a\sum_i i + b\sum_i i^2 \end{cases}$$

donde N es el número de años de los que tenemos datos en la tabla (en este caso, 4 años que corresponden a las 4 columnas) y los coeficientes a y b son los coeficientes de la recta de regresión que denominamos recta de tendencia, la fórmula de la cual es $T_i = a + b$, donde i hace referencia al año que se indica en la escala de la tabla superior.

$$\begin{cases} \sum_i \overline{y}_i = Na + b\sum_i i \\ \sum_i \overline{y}_i \cdot i = a\sum_i i + b\sum_i i^2 \end{cases} \Rightarrow \begin{cases} 73 = 4a + 2b \\ 23 = 2a + 6b \end{cases}$$

Resolveremos el sistema por el método que consideremos más adecuado (sería muy fácil por reducción) y encontramos los valores a = 19,6 y b = –2,7, con los cuales se concluye que T_i = 19,6 – 2,7i.

Para interpretar esta componente de la serie, T_i, que se llama tendencia y que nos explica el comportamiento del fenómeno a «largo plazo», nos fijamos en el valor de la pendiente b = 2,7. Que por ser negativa nos permite explicar que los gastos del departamento van decreciendo año tras año, y además podemos detallar que el valor de la media de gastos cuatrimestrales del departamento, $\overline{y}_i$, cada año ha disminuido en 2,7 miles de euros.

Cálculo de la componente estacional

La segunda componente de la serie que hay que calcular es la componente estacional, que denotaremos por e_k, que nos permitirá analizar el comportamiento del fenómeno por períodos dentro del año (en nuestro ejercicio, por cuatrimestres), valorando en cuáles de ellos los valores están por encima o por debajo de un valor global que denotaremos por M y que llamaremos media global corregida.

Los valores de e_k estarán expresados en valores absolutos y en las mismas unidades que los datos de la tabla original (en este ejercicio en miles de euros).

Para abordar los cálculos de e_k trabajaremos ampliando la tabla en diferentes columnas hacia la derecha, ya que nos interesa hacer un trabajo por períodos, en este caso, por cuatrimestres.

En la columna primera calcularemos la media de los valores de cada cuatrimestre y esta media la denotaremos por $\overline{y}_k = \frac{\sum_i y_{ij}}{N}$.

	2008	2009	2010	2011		$\overline{y}_k$	$\overline{y}'_k$	e_k	I_k
1.^er cuatrimestre	26	25	21	20		23	23	3,85	120,10 %
2.º cuatrimestre	18	15	12	10		13,75	14,65	-4,5	76,50 %
3.^er cuatrimestre	22	20	18	12	TOTALES ↓	18	19,8	0,65	103,39 %
							M = 19,15		

En la siguiente columna calcularemos las medias corregidas $\overline{y}'_k$ (donde eliminamos en cada dato el valor proporcional a la tendencia que le podemos asignar a cada período), suponiendo que este decrecimiento *b* = –2,7 ha sido constante a lo largo de los períodos del año.

En este ejercicio, *b/m* = –2,7 / 3 = – 0,9. Al ser una tendencia negativa, los valores de las medias corregidas $\overline{y}'_k$ seran mayores que las medias originales $\overline{y}_k$, ya que estas se calculan así:

$$\overline{y}'_k = \overline{y}_k - \frac{b}{m}(k - 1)$$

Así:

$$\overline{y}'_1 = \overline{y}_1 - \frac{b}{m}(1-1) = 23 - (-0,9)(1-1) = 23$$

$$\overline{y}'_2 = \overline{y}_2 - \frac{b}{m}(2-1) = 13,75 - (-0,9)(2-1) = 14,65$$

$$\overline{y}'_3 = \overline{y}_3 - \frac{b}{m}(3-1) = 18 - (-0,9)(3-1) = 19,8$$

A continuación, en una celda inferior calculamos *M*, la media de las medias corregidas, ya que:

$$M = \frac{\sum_k \overline{y}'_k}{m} = \frac{23 + 14,65 + 19.8}{3} = 19,15$$

y representa el valor de referencia para analizar los valores de los períodos, mediante $e_k = \overline{y}'_k - M$, que calcularemos en la siguiente columna:

$$e_1 = \overline{y}'_1 - M = 23 - 19,15 = 3,85$$

$$e_2 = \overline{y}'_2 - M = 14,65 - 19,15 = -4,5$$

$$e_3 = \overline{y}'_3 - M = 19,8 - 19,15 = 0,65$$

La componente estacional e_k nos explicita el comportamiento por períodos en valores absolutos, indicándonos signo y cantidad en las mismas unidades que los datos originales (miles de euros). Así, diremos que durante el primer cuatrimestre los gastos del departamento de nuestra empresa tienen un valor por encima de la la media de 3850 euros, mientras que los gastos del segundo cuatrimestre son de 4500 euros por debajo de la media y el tercer cuatrimestre son superiores solo en 650 euros a dicha media. El valor de la media de referencia sería M = 19.150 euros que podría representar un promedio de gastos cuatrimestrales global.

Otra interpretación de estos datos se puede dar con los índices estacionales $I_k = \frac{\overline{y}_k^{'}}{M} \cdot 100$ que calculamos en la siguiente columna, donde se indica en forma de porcentaje (valor relativo, donde M representa el 100 %) la misma información que la componente estacional, pero que al tener carácter porcentual es más fácil de presentar sin particularizar y dar el valor de M.

$$I_1 = \frac{\overline{y}_1^{'}}{M} \cdot 100 = \frac{23}{19{,}15} \cdot 100 = 120{,}10\ \%$$

$$I_2 = \frac{\overline{y}_2^{'}}{M} \cdot 100 = \frac{14{,}65}{19{,}15} \cdot 100 = 76{,}50\ \%$$

$$I_3 = \frac{\overline{y}_3^{'}}{M} \cdot 100 = \frac{19{,}8}{19{,}15} \cdot 100 = 103{,}39\ \%$$

Esta componente estacional expresada en términos de porcentaje, los índices estacionales, nos permite analizar e interpretar el comportamiento de los datos de la serie, es decir, los valores de los gastos del departamento por cuatrimestres.

Podremos afirmar que los gastos eran mayores en el primer cuatrimestre, con valores un 20,10 % superiores a la media anual, mientras que los valores del tercer cuatrimestre siquiera superan dicho valor en un 3,39 %. Cabe destacar que los gastos disminuyen el segundo cuatrimestre con valores que están alrededor del 76,50 % del valor de dicha media anual global, que podríamos considerar el valor de M = 19,15 miles de euros.

Cálculo de la componente residual o errática

En tercer lugar, hay que calcular la componente errática o residual que denotamos por r_{ik}, que nos permite destacar el comportamiento de algún dato y_{ik}, el valor del cual no se pueda explicar por las anteriores componentes, lo que permitirá inferir que por cualquier causa por identificar (motivos extraordinarios) este valor no está dentro del patrón de comportamiento que hemos encontrado y con el que hemos interpretado los datos originales para explicar el fenómeno. Los valores de esta componente errática deben ser pequeños en valor, variados en signo y sin regularidad ni patrón. Nos permiten ver que los datos reales no se ajustan completamente

al patrón que hemos encontrado con la tendencia y la componente estacional. Es por eso que si algún valor es muy alto o bajo dejará identificar algún dato que corresponda al período de un año, el valor del cual se aleja mucho de los valores que cabría esperar.

La calcularemos así $r_{ik} = y_{ik} - T_i - e_k$ en cada celda de la tabla. Para tal fin, primeramente hay que calcular la tendencia para cada año de la tabla $T_i = 19,6 - 2,7i$.

$$T_{2008} = T_{-1} = 19,6 - 2,7(-1) = 22,3$$
$$T_{2009} = T_0 = 19,6 - 2,7 \cdot 0 = 19,6$$
$$T_{2010} = T_1 = 19,6 - 2,7 \cdot 1 = 16,9$$
$$T_{2011} = T_2 = 19,6 - 2,7 \cdot 2 = 14,2$$

Los valores de la componente estacional están en columna de la tabla e_k:

$e_1 = 3,85$ $\qquad$ $e_2 = -4,5$ $\qquad$ $e_3 = 0,65$

Así, los cálculos para cada celda la están en la siguiente tabla:

	2008	2009	2010	2011
1.er cuatrimestre	26 – 22,3 – 3,85 = –0,15	25 – 19,6 – 3,85 = 1,55	21 – 16,9 – 3,85 = 0,25	20 – 14,2 – 3,85 = 1,95
2.o cuatrimestre	18 – 22,3 + 4,5= 0,2	15 – 19,6 + 4,5 = –0,1	12 – 16,9 + 4,5 = –0,4	10 – 14,2 + 4,5 = 0,3
3.er cuatrimestre	22 – 22,3 – 0,65 = –0,95	20 – 19,6 –0,65 = –0,25	18 – 16,9 – 0,65= 0,45	12 – 14,2 – 0,65 = –2,85

De estos resultados, los valores que nos llaman más la atención serían los correspondientes al primer cuatrimestre de 2011, que es superior a lo que cabría esperar con $r_{2011,1} = 1,95$, y el tercer cuatrimestre del mismo año con un valor aún más inferior con $r_{2011,3} = -2,85$. Habría que estudiar si alguna circunstancia extraordinaria justifica estos valores o, por el contrario, indica que el comportamiento del fenómeno está cambiando sustancialmente. Habría que estar atentos a los próximos datos del año siguiente.

b) Estima los valores de los gastos que se pueden esperar para el año 2012.

Para hacer las estimaciones que nos piden en este apartado, recordemos que no podemos prever la componente residual, por lo que calcularemos:

$$y_{ik} = T_i + e_k$$

Considerando la escala que hemos adoptado en la tabla para los años, con el propósito de simplificar los cálculos de la tendencia, podemos establecer que si $2009 \Rightarrow i = 0$, esto comporta $2012 \Rightarrow i = 3$, lo que nos permitirá obtener el valor

de la tendencia para este año: $T_{2012} = T_3 = 19,6 - 2,7 \cdot 3 = 11,5$ y como conocemos la componente estacional $e_1 = 3,85$, $e_2 = -4,5$ y $e_3 = 0,65$, podremos estimar los valores de los gastos del año 2012 por quadrimestres.

$y_{2012,1qua} = 11,5 + 3,85 = 15,35$ miles de euros

$y_{2012,2qua} = 11,5 - 4,5 = 7$ miles de euros

$y_{2012,3qua} = 11,5 + 0,65 = 12,15$ miles de euros

Ejercicio 2

Los siguientes datos, extraídos del INI, nos muestran las pernoctaciones hoteleras en la Comunidad Valenciana, desde el año 2006 hasta el 2010.

Realiza un análisis del fenómeno, obteniendo las diferentes componentes de la serie temporal para el ajuste analítico, suponiendo un modelo aditivo. Interpreta el significado de cada una de las componentes.

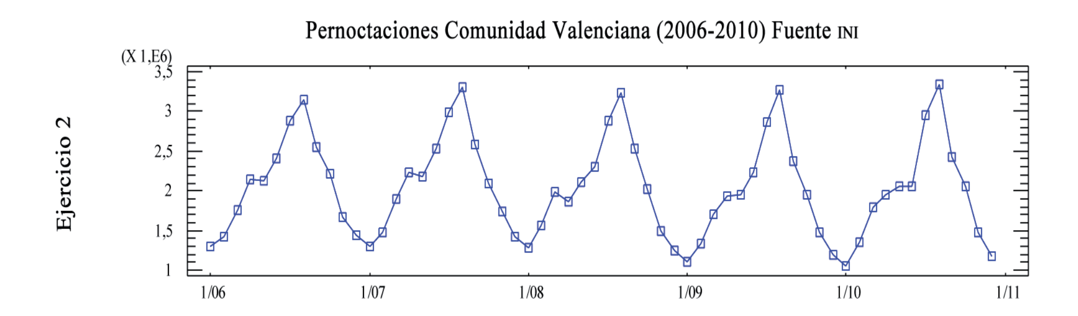
Haz las previsiones que podemos esperar para los años 2012 y 2013.

	2006	2007	2008	2009	2010
Enero	1296648	1307384	1289627	1115865	1061450
Febrero	1424453	1472806	1562471	1329013	1347885
Marzo	1750282	1892925	1993175	1700733	1800323
Abril	2152783	2226734	1868049	1925454	1943080
Mayo	2131194	2181889	2114631	1952409	2049616
Junio	2399782	2523555	2311444	2241147	2049616
Julio	2884491	2983227	2877028	2871011	2945899
Agosto	3153407	3308833	3227055	3274561	3341961
Septiembre	2540711	2580090	2524888	2381776	2424567
Octubre	2207203	2097005	2012838	1954615	2053500
Noviembre	1674433	1741296	1498967	1471579	1472610
Diciembre	1437036	1420988	1251809	1201342	1174451

Solución

Análisis del modelo y gráfica

Para empezar, hay que observar la gráfica de los datos y comprobar que pueden ser analizados por el método del ajuste analítico, suponiendo un modelo aditivo. Es fácil observar un patrón que se repite con bastante regularidad año tras año con un marcado comportamiento estacional reiterado a lo largo de la colección de datos que presentamos. En cambio, no se observa una tendencia de marcada pendiente, ya que los datos se mantienen bastante constantes a largo plazo.



Cálculo de la tendencia

Consideraremos los datos de la serie siguiente que hace referencia a las pernoctaciones mensuales hoteleras en la Comunidad Valenciana en los últimos cinco años:

	2006	2007	2008	2009	2010	
Enero	1296648	1307384	1289627	1115865	1061450	
Febrero	1424453	1472806	1562471	1329013	1347885	
Marzo	1750282	1892925	1993175	1700733	1800323	
Abril	2152783	2226734	1868049	1925454	1943080	
Mayo	2131194	2181889	2114631	1952409	2049616	
Junio	2399782	2523555	2311444	2241147	2049616	
Julio	2884491	2983227	2877028	2871011	2945899	
Agosto	3153407	3308833	3227055	3274561	3341961	
Septiembre	2540711	2580090	2524888	2381776	2424567	
Octubre	2207203	2097005	2012838	1954615	2053500	
Noviembre	1674433	1741296	1498967	1471579	1472610	

Diciembre	1437036	1420988	1251809	1201342	1174451	TOTALES
$\overline{y}_i$	2087702	2144728	2044332	1951625	1972080	10200467
i	–2	–1	0	1	2	0
$\overline{y}_i \cdot i$	–4175404	–2144728	0	1951625	3944160	–424346
i^2	4	1	0	1	4	10

Podemos ver que en la parte inferior de la tabla hemos calculado las medias mensuales de cada año $\overline{y}_i$, los valores de la escala i y las dos files inferiores $\overline{y}_i \cdot i,\ i^2$, que nos permitirá calcular en la columna *totales* la suma de los valores de cada fila. En todos los cálculos hemos redondeado a números enteros por tratarse de número de pernoctaciones.

Con estos datos, plantearemos el siguiente sistema:

$$\begin{cases} \sum_i \overline{y}_i = Na + b\sum_i i \\ \sum_i \overline{y}_i \cdot i = a\sum_i i + b\sum_i i^2 \end{cases} \Rightarrow \begin{cases} 10200467 = 5a + 0b \\ -424346 = 0a + 10b \end{cases}$$

donde los coeficientes a y b son los coeficientes de la recta de regresión que denominamos recta de tendencia, la formula es $T_i = a + bi$, donde i hace referencia al año que se indica en la escala de la tabla superior.

Podemos ver en este ejemplo que la estrategia de darle el valor $i = 0$ al año correspondiente en la columna central, cuando N (número de años) es impar, simplifica mucho la resolucion del sistema. Así:

$$a = \frac{10200467}{5} = 2040093{,}4 \qquad b = \frac{-424346}{10} = -42434{,}6$$

son los coeficientes de $T_i = a + bi = 2040093{,}4 - 42434{,}6i$ que es la ecuación de la recta de tendencia y nos permite interpretar que a largo plazo, a partir de estos datos, la media mensual del número de pernoctaciones disminuye en 42.435 pernoctaciones por año. Por tratarse de una cantidad pequeña en relación a los datos de la serie, es la razón por la cual en la gráfica no observábamos claramente el decrecimiento.

Cálculo de la componente estacional

Para calcular la componente estacional, trabajaremos las columnas que se pueden ver a la derecha de la tabla correspondiente a la serie de pernoctaciones de la Comunidad Valenciana (2006-2010).

Recordemos que en la primera columna, $\overline{y}_k = \frac{\sum_i y_{ij}}{N}$ son las medias de las pernoctaciones de cada mes calculadas con los datos originales, por filas.

La columna siguiente corresponde a las medias corregidas, la fórmula de las cuales és $\overline{y}_k^{'} = \overline{y}_k - \frac{b}{m}(k-1)$ donde $b/m = –42434,6/12 = –3536,2$; recordemos que $m = 12$ porque, al tratarse de datos mensuales, tenemos 12 datos por año.

M es la media global corregida y es la media de las medias corregidas:

$$M = \frac{\sum_k \overline{y}_k^{'}}{m}$$

que podemos interpretar como el valor medio de las pernoctaciones mensuales y que representará el 100 % o valor de referencia.

Después, las dos columnas siguientes muestran los valores de la componente estacional $e_k = \overline{y}_k^{'} - M$ y los índices estacionales $I_k = \frac{\overline{y}_k^{'}}{M} \cdot 100$, donde podemos ver en qué meses el número de pernoctaciones está por encima o por debajo del valor de M. Así pues, e_k nos presenta esta desviación en cantidades absolutas (número de pernoctaciones) mientras que I_k lo indica de manera porcentual.

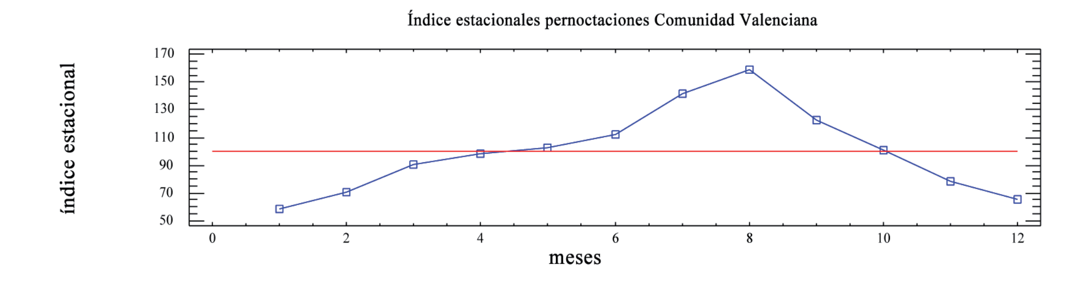
Podemos interpretar que el mes con más pernoctaciones es agosto con un valor por encima de la media es 59,55 %, mientras que el mes que registra menos pernoctaciones es enero, que registra tan solo unos valores que corresponden al 58,95 % del valor de referencia M. Podemos observar también que los valores más importantes se concentran en los meses de agosto, julio y septiembre, mientras que los meses de menor afluencia corresponden a enero, diciembre y febrero.

Podemos observar un comportamiento estacional muy marcado.

Tabla correspondiente a la serie de pernoctaciones en la Comunidad Valenciana (2006-2010)

	2006	2007	2008	2009	2010		$\overline{y}_k$	$\overline{y}_k^{'}$	e_k	I_k
Enero	1296648	1307384	1289627	1115865	1061450		1214195	1214195	–845348	58,95
Febrero	1424453	1472806	1562471	1329013	1347885		1427326	1430862	–628681	69,47
Marzo	1750282	1892925	1993175	1700733	1800323		1827488	1834560	–224982	89,08
Abril	2152783	2226734	1868049	1925454	1943080		2023220	2033828,6	–25714	98,75
Mayo	2131194	2181889	2114631	1952409	2049616		2085948	2100092,6	40550	101,97
Junio	2399782	2523555	2311444	2241147	2049616		2305109	2322789,8	263247	112,78
Julio	2884491	2983227	2877028	2871011	2945899		2912331	2933548,4	874006	142,44
Agosto	3153407	3308833	3227055	3274561	3341961		3261163	3285916,8	1226374	159,55
Septiembre	2540711	2580090	2524888	2381776	2424567		2490406	2518696	459154	122,29
Octubre	2207203	2097005	2012838	1954615	2053500		2065032	2096858	37316	101,81
Noviembre	1674433	1741296	1498967	1471579	1472610		1571777	1607139	–452403	78,03
Diciembre	1437036	1420988	1251809	1201342	1174451	**TOTALES**	1297125	1336023,4	–723519	64,87
$\overline{y}_i$	2087702	2144728	2044332	1951625	1972080	10200467		**M** = 2059542		
i	–2	–1	0	1	2	0				
$\overline{y}_i \cdot i$	–4175404	–2144728	0	1951625	3944160	-424346				
i^2	4	1	0	1	4	10				
Tendència	2124962,6	2082528	2040093	1997658,8	1955224,2		b/m=- 42434,6/12= -3536,2			

Recta de tendencia: $T_i = a + bi = 2040093,4 - 42434,6i$ con $i = 0$ que corresponde al año 2008, como podemos ver en la tabla.



Cálculo de la componente residual o errática

Para calcular la componente residual $r_{ik} = y_{ik} - T_i - e_k$ habrá que calcular primero el valor de la tendencia para cada año, y consideraremos para la componente estacional e_k los valores que podemos encontrar en la tabla ya calculados. Así:

$$T_{2006} = 2040093,4 - 42434,6(-2) = 2124962,6$$

$T_{2007} = 2040093,4 - 42434,6(-1) = 2082528$
$T_{2008} = 2040093,4 - 42434,6 \cdot 0 = 2040093,4$
$T_{2009} = 2040093,4 - 42434,6 \cdot 1 = 1997658,8$
$T_{2010} = 2040093,4 - 42434,6 \cdot 2 = 1955224,2$

y consideraremos los valores de la componente estacional:

$e_1 = -845.348$ $e_2 = -628.681$ $e_3 = -224.982$
$e_4 = -25.714$ $e_5 = 40.550$ $e_6 = 263.247$
$e_7 = 874.00$ $e_8 = 1.226.374$ $e_9 = 459.154$
$e_{10} = 37.316$ $e_{11} = -452.403$ $e_{12} = -723.519$

Con estos datos, los valores de la componente residual están calculados en la siguiente tabla (para ver mejor los datos a efectos de interpretación hemos redondeado los resultados en número exacto de pernoctaciones y hemos reseñado con rojo los valores negativos).

Por ser la componente residual estas cifras podrían sorprendernos, pero hay que considerar que si comparamos los valores más extremos resaltados (que posteriormente comentaremos), estamos hablando de magnitudes cercanas a +/– 150.000 frente al valor global de la tabla $M = 2.059.542$. Por lo cual, estamos hablando de valores relativos cercanos al 7 %.

Hacemos esta advertencia porque todos sabemos que las cantidades que hay que obtener de la componente residual deberían ser pequeñas en valor absoluto, y que no presentan ninguna regularidad. Ya sabemos que estamos calculando las cantidades no explicadas por nuestro modelo y que permitirán resaltar aquellos valores puntuales, que por razones no predecibles, muestran divergencia del valor que cabría esperar, atendiendo a las componentes de la tendencia y estacional.

Como se puede ver, en marzo y agosto de 2006, los valores estuvieron por debajo las previsiones, mientras que en abril y junio de 2007 los valores estuvieron por encima. Hay que fijarse con el comportamiento de los datos en los meses de febrero, marzo y abril de 2008 y junio, julio y agosto de 2010 que presentan, en ambos casos, la secuencia de tres meses con valores que parece que se «compensan» en la misma temporada.

	residual 2006	**residual 2007**	**residual 2008**	**residual 2009**	**residual 2010**
Enero	17033	70204	94881	–36446	–48427
Febrero	–71829	18959	**151058**	–39965	21341
Marzo	**–149698**	35379	**178064**	–71943	70081
Abril	53534	**169920**	**–146331**	–46491	13570
Mayo	–34319	58811	33987	–85800	53842
Junio	11572	**177780**	8103	–19759	**–168856**
Julio	–114478	26693	–37071	–654	**116669**

Agosto	**–197930**	–69	–39413	50528	**160362**
Septiembre	–43405	38408	25641	–75036	10189
Octubre	44925	–22839	–64571	–80359	60960
Noviembre	1874	111171	–88723	–73676	–30211
Diciembre	35592	61979	–64765	–72798	–57254

Previsiones para los años 2012 y 2013

Hacer predicciones para los años posteriores implica que el análisis de nuestro modelo sea vigente y que ninguna otra circunstancia ajena altere las regularidades que hemos reseñado con nuestro modelo (la tendencia ligeramente decreciente y el comportamiento mensual ya comentado).

Si queremos prever las cantidades de los años 2012 y 2013 será necesario calcular la tendencia sustituyendo los valores $i = 4$ (para el año 2012) e $i = 5$ (para el año 2013). Estos valores serían los que corresponderían a estos años en la escala de los valores i de la primera parte de la tabla. Así:

$T_{2012} = 2040093{,}4 - 42434{,}6 \cdot 4 = 1870355$
$T_{2013} = 2040093{,}4 - 42434{,}6 \cdot 5 = 1827920{,}4$

y con la componente estacional antes mencionada, podremos hacer las previsiones mensuales para los años 2012 y 2013, estimando $y_{ik} = T_i + e_k$.

	Previsiones 2012	**Previsiones 2013**
Enero	1870355 – 845348 = 1025007	1827920,4 – 845348 = 982573
Febrero	1870355 – 628681 = 1241674	1827920,4 – 628681 = 1199240
Marzo	1870355 – 224982 = 1645373	1827920,4 – 224982 = 1602938
Abril	1870355 – 25714 = 1844641	1827920,4 – 25714 = 1802207
Mayo	1870355 + 40550 = 1910905	1827920,4 + 40550 = 1868471
Junio	1870355 + 263247 = 2133602	1827920,4 + 263247 = 2091168
Julio	1870355 + 874006 = 2744361	1827920,4 + 874006 = 2701926
Agosto	1870355 + 1226374 = 3096729	1827920,4 + 1226374 = 3054295
Septiembre	1870355 + 459154 = 2329509	1827920,4 + 459154 = 2287074
Octubre	1870355 + 37316 = 1907671	1827920,4 + 37316 = 1865236
Noviembre	1870355 – 452403 = 1417952	1827920,4 – 452403 = 1375517
Diciembre	1870355 – 723519 = 1146836	1827920,4 – 723519 = 1104401

Ejercicio 3

Con los siguientes datos, extraídos de la DGT, que nos muestran las nuevas licencias de todos los tipo de carnés de conducir en la Comunidad Valenciana, desde el año 2008 hasta el 2010, realiza un análisis del fenómeno, obteniendo las diferentes componentes de la serie temporal por el ajuste analítico.

Interpreta el significado de cada una de las componentes.

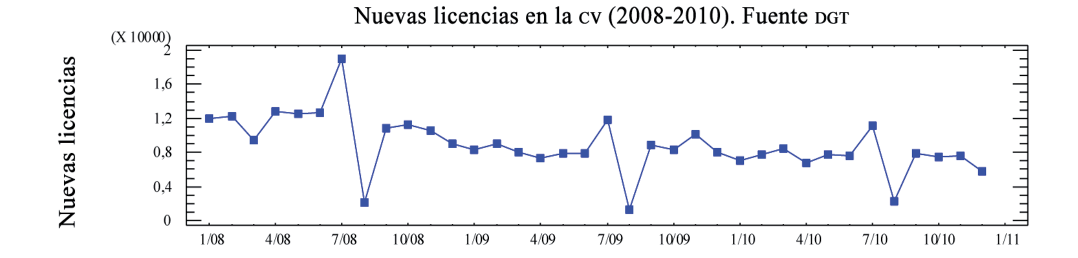
Haz las previsiones que podemos esperar para los años 2011 y 2012.

	2008	2009	2010
Enero	12031	8380	7071
Febrero	12208	8993	7685
Marzo	9497	7973	8444
Abril	12862	7360	6781
Mayo	12567	7874	7728
Junio	12723	7881	7585
Julio	19003	11820	11138
Agosto	2147	1346	2205
Septiembre	10826	8876	7901
Octubre	11196	8374	7427
Noviembre	10628	10137	7665
Diciembre	9064	8083	5746

Solución

Análisis del modelo y gráfico

Para empezar, hay que observar la gráfica de los datos y comprobar que pueden ser analizados por el método del ajuste analítico, suponiendo un modelo aditivo.



Podemos observar una tendencia ligeramente decreciente y un comportamiento estacional bastante constante, exceptuando los meses de julio y agosto.

Cálculo de la tendencia

Consideraremos los datos de la serie siguiente que hace referencia al número de las nuevas licencias de la totalidad de los tipos de permisos de conducir en la Comunidad Valenciana, en los años 2008 al 2010:

	2008	2009	2010	
Enero	12031	8380	7071	
Febrero	12208	8993	7685	
Marzo	9497	7973	8444	
Abril	12862	7360	6781	
Mayo	12567	7874	7728	
Junio	12723	7881	7585	
Julio	19003	11820	11138	
Agosto	2147	1346	2205	
Septiembre	10826	8876	7901	
Octubre	11196	8374	7427	
Noviembre	10628	10137	7665	
Diciembre	9064	8083	5746	TOTALES
$\overline{y}_i$	11229	8091	7281	26602
i	-1	0	1	0
$\overline{y}_i \cdot i$	-11229	0	7281	-3948
i^2	1	0	1	2

Podemos ver que en la parte inferior de la tabla hemos calculado las medias mensuales de cada año $\overline{y}_i$, los valores de la escala i y las dos filas inferiores $\overline{y}_i \cdot i$, i^2, que nos permitirá calcular en la columna *Totales* la suma de los valores de cada fila.

Con estos datos, plantearemos el siguiente sistema:

$$\begin{cases} \sum_i \overline{y}_i = Na + b\sum_i i \\ \sum_i \overline{y}_i \cdot i = a\sum_i i + b\sum_i i^2 \end{cases} \Rightarrow \begin{cases} 26602 = 3a + 0b \\ -3948 = 0a + 2b \end{cases}$$

donde a y b son los coeficientes de la recta de regresión que denominamos recta de tendencia, la formula es $T_i = a + b$, donde i hace referencia al año que se indica en la escala de la tabla superior.

Podemos ver en este ejemplo que la estrategia de darle el valor $i = 0$ al año correspondiente en la columna central, cuando N (número de años) es impar simplifica mucho la resolucion del sistema. Así:

$$a = \frac{26602}{3} = 8867{,}36 \qquad\qquad b = \frac{-3948}{2} = -1974$$

son los coeficientes de $T_i = a + bi = 8867{,}36 - 1974i$ que es la ecuación de la recta de tendencia y nos permite interpretar que a largo plazo, a partir de estos datos, la media mensual del número de nuevas licencias disminuye en 1974 nuevos permisos por año. Por tratarse de una cantidad pequeña en relación a los datos de la serie, es la razón por la cual en la gráfica no observábamos claramente el decrecimiento.

Cálculo de la componente estacional

Para calcular la componente estacional trabajaremos las columnas que se pueden ver a la derecha de la tabla siguiente.

Recordemos que en la primera columna, $\overline{y}_k = \frac{\sum_i y_{ij}}{N}$ son las medias de los datos de cada mes, calculadas con los datos originales, por filas.

La columna siguiente corresponde a las medias corregidas, la fórmula de las cuales es $\overline{y}_k^{'} = \overline{y}_k - \frac{b}{m}(k-1)$ donde $b/m = -1974/12 = -164{,}5$. Recordemos que $m = 12$ porque, al tratarse de datos mensuales, tenemos 12 datos por año.

M es la media global corregida y es la media de las medias corregidas.

$$M = \frac{\sum_k \overline{y}_k^{'}}{m}$$

que podemos interpretar como el valor medio de las nuevas medias corregidas mensuales y que representará el 100 % o valor de referencia.

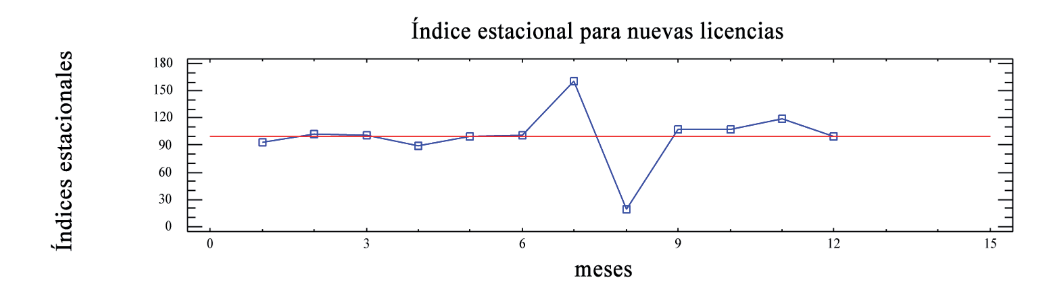
Después, las dos columnas siguientes muestran los valores de la componente estacional $e_k = \overline{y}_k' - M$ y los índices estacionales $I_k = \frac{\overline{y}_k}{M} \cdot 100$, donde podemos ver en qué meses el número de nuevas licencias está por encima o debajo del valor de *M*. Así pues, e_k nos presenta esta desviación en cantidades absolutas (número de pernoctaciones) mientras que I_k lo indica de manera porcentual.

Podemos interpretar que el mes con menos licencias nuevas expedidas es agosto con un valor muy por debajo de la media del 31 %, mientras que el mes que registra más licencias expedidas es julio, que registra un valor del 53 % por encima del valor de referencia *M*. Podemos observar también que el resto de meses tienen valores bastante cercanos al 100 %, lo que podemos interpretar como que este fenómeno no tiene un comportamiento estacional marcado, excepto los meses de julio y agosto antes citados.

Tabla correspondiente a la serie de nuevas licencias expedidas de todos los tipos de carnés en la Comunidad Valenciana (2008-2010)

	2008	2009	2010			$\overline{y}_k$	$\overline{y}_k'$	e_k	I_k
Enero	12031	8380	7071			9161	9.161	–611	93,74
Febrero	12208	8993	7685			9629	9.793	21	100,22
Marzo	9497	7973	8444			8638	8967	–805	91,76
Abril	12862	7360	6781			9001	9494,5	–278	97,16
Mayo	12567	7874	7728			9390	10047,6667	276	102,82
Junio	12723	7881	7585			9396	10218,8333	447	104,57
Julio	19003	11820	11138			13987	14974	5202	153,23
Agosto	2147	1346	2205			1899	3050,83333	–6721	31,22
Septiembre	10826	8876	7901			9201	10517	745	107,62
Octubre	11196	8374	7427			8999	10479,5	707	107,24
Noviembre	10628	10137	7665			9477	11121,6667	1350	113,81
Diciembre	9064	8083	5746	TOTALES		7631	9440,5	–332	96,61
$\overline{y}_i$	11229	8091	7281	26602		M =	9.772		
i	–1	0	1	0					
$\overline{y}_i \times i$	–11229	0	7281	–3948					
i^2	1	0	1	2					
Tendencia	10841,36111	8867,361	6893,361111						
					b/m = –164,5				

Recta de tendencia: $T_i = a + bi = 8867{,}36 - 1974i$ con $i = 0$ que corresponde al año 2009, como se pueden ver en la tabla.



Cálculo de la componente residual o errática

Para hallar la componente residual $r_{ik} = y_{ik} - T_i - e_k$ se deberá calcular primero el valor de la tendencia para cada año, y consideraremos para la componente estacional e_k los valores que podemos encontrar en la tabla ya calculados. Así:

$T_{2008} = 8867{,}36 - 1974(-1) = 10841{,}36$
$T_{2009} = 8867{,}36 - 1974 \cdot 0 = 8867{,}36$
$T_{2010} = 8867{,}36 - 1974 \cdot 1 = 6893{,}36$

y consideraremos los valores de la componente estacional:

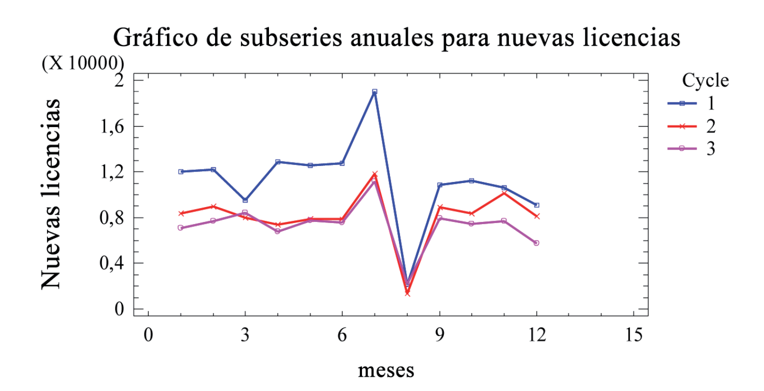
$e_1 = -611$	$e_2 = 21$	$e_3 = -805$
$e_4 = -278$	$e_5 = 276$	$e_6 = 447$
$e_7 = 5.202$	$e_8 = -6.721$	$e_9 = 745$
$e_{10} = 707$	$e_{11} = 1.350$	$e_{12} = -332$

Con estos datos, los valores de la componente residual están calculados en la siguiente tabla (para ver mejor los datos a efectos de interpretación hemos redondeado los resultados en número exacto de licencias y hemos reseñado con rojo los valores negativos)

Hacemos esta advertencia porque todos sabemos que las cantidades que hay que obtener de la componente residual deberían ser pequeñas en valor absoluto, y que no presentan ninguna regularidad. Ya sabemos que estamos calculando las cantidades no explicadas por nuestro modelo y que permitirán resaltar aquellos valores puntuales que por razones no predecibles, muestran divergencia del valor que cabría esperar, atendiendo a las componentes de la tendencia y estacional.

Como se puede ver, en la tabla siguiente hemos remarcado en rojo los datos que están por debajo de los valores que cabría esperar y con azul los valores que están por encima. Para interpretar conjuntamente los datos reales y los resultados de la componente estacional hemos introducido debajo de la tabla un gráfico de las subseries anuales de los datos.

	Residual 2008	Residual 2009	Residual 2010
Enero	1801	124	789
Febrero	1346	105	771
Marzo	–539	–89	2356
Abril	2298	–1230	165
Mayo	1450	–1269	559
Junio	1435	–1433	245
Julio	2960	–2249	–957
Agosto	–1973	–800	2033
Septiembre	–760	–736	263
Octubre	–353	–1201	–174
Noviembre	–1563	–80	–578
Diciembre	–1446	–453	–816



Podemos observar que los valores de los datos de la componente residual del año 2008, efectivamente, corresponden a los meses, los datos de los cuales se alejan bastante del comportamiento que corresponde al patrón de los otros años. Los datos de agosto distorsionan un poco porque no siguen el comportamiento de la tendencia anual; es como si dijéramos que es un mínimo porque paran los servicios y es independiente de los valores de la tendencia.

La mayor parte de los datos reseñados corresponden al año 2008 porque se puede observar mejor en este gráfico que no sigue el mismo patrón estacional que los años siguientes, si exceptuamos los valores de los meses de verano.

En todo caso, cabe destacar que es una serie demasiado corta (tres años) para hacer un análisis muy riguroso.

Previsiones para los años 2012 y 2013

Hacer predicciones para los años posteriores que implica que el análisis de nuestro modelo sea vigente y que ninguna otra circunstancia ajena altere las regularidades que hemos reseñado con nuestro modelo (la tendencia ligeramente decreciente y el comportamiento mensual ya comentado).

Si queremos prever las cantidades de los años 2012 y 2013 será necesario calcular la tendencia sustituyendo los valores $i = 2$ (para el año 2011) e $i = 3$ (para el año 2012). Estos valores serían los que corresponderían a estos años en la escala de los valores i de la primera parte de la tabla. Así:

$$T_{2011} = 8867{,}36 - 1974 \cdot 2 = 4919{,}36$$
$$T_{2012} = 8867{,}36 - 1974 \cdot 3 = 2945{,}36$$

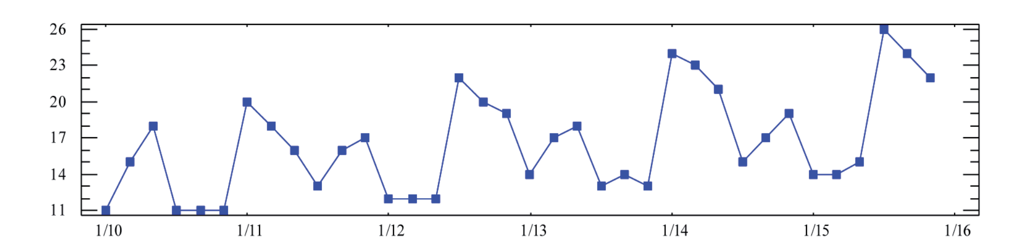
y con la componente estacional antes mencionada, podremos hacer las previsiones mensuales para los años 2011 y 2012, estimando $y_{ik} = T_i + e_k$.

	Previsiones 2011	Previsiones 2012
Enero	4919,36 – 611 = 4308	2945,36 – 611 = 2334
Febrero	4919,36 + 21 = 4940	2945,36 + 21 = 2966
Marzo	4919,36 – 805 = 4114	2945,36 – 805 = 2140
Abril	4919,36 – 278 = 4642	2945,36 – 278 = 2668
Mayo	4919,36 + 276 = 5195	2945,36 + 276 = 3221
Junio	4919,36 + 447 = 5366	2945,36 + 447 = 3292
Julio	4919,36 + 5202 = 10.121	52945,36 + 202 = 8147
Agosto	4919,36 – 6721 = –1802	2945,36 – 6.721 = –3776
Septiembre	4919,36 + 745 = 5664	2945,36 + 745 = 3690
Octubre	4919,36 + 707 = 5627	2945,36 + 707 = 3653
Noviembre	4919,36 + 1350 = 6269	2945,36 + 1.350 = 4295
Diciembre	4919,36 – 332 = 4588	2945,36 – 332 = 2614

Hay que hacer notar que los resultados de la tabla anterior están redondeados por tratarse de número de nuevas licencias que podemos prever para los años que se señalan. No se debe olvidar, insistimos, que la fiabilidad de estas previsiones está en función de la hipótesis de que el fenómeno mantenga el comportamiento decreciente que nos ha indicando la tendencia.

Ejercicio 4

La siguiente gráfica es la representación de una serie temporal donde se detallan los datos bimensuales de 6 años. Si tuviéramos que calcular la tendencia y la componente estacional de dicha serie por el método de las medias móviles, explica la elección del número de datos que habría que considerar para el cálculo de las medias (p) en cada caso, justificando la respuesta.



Solución

Para empezar la resolucion por el método de las medias móviles hay que elegir el número adecuado de observaciones por incluir en cada media. Este número depende de la componente que queremos encontrar.

Comenzaremos por explicar este número en el caso de la tendencia.

Este número que denotaremos por p es el mínimo común múltiplo, MCM en adelante, el número de observaciones por año (en este caso 6 para ser observaciones bimensuales, es decir, tenemos un dato cada dos meses, por lo que tenemos 6 datos por año) y el número de observaciones que incluye cada «período» de la gráfica (9 en este caso porque tal y como se ve en la gráfica, la serie tiene un comportamiento o patrón que se repite cada 9 puntos aproximadamente).

Así pues, para calcular la tendencia el número de observaciones por tomar para cada media móvil es p, donde

$$p = MCM\,(6{,}9) = 18$$

Esta elección nos asegura que al hacer el suavizado de 18 observaciones, todas las fluctuaciones anuales y estacionales se contemplan en cada medio, y así sustituimos cada valor que se abandona en una media por otro que sería equivalente a la media siguiente a fin de conseguir la tendencia, que explica el comportamiento del fenómeno a largo plazo de la serie.

Estudiamos ahora cuál sería el número de observaciones a considerar en el caso de la componente estacional.

En este caso, el número p es el número de observaciones que hay que tomar para tener un año completo. Es decir, en nuestro problema, $p = 6$.

En esta elección, cada dato es sustituido en un promedio por otra que tiene el mismo comportamiento estacional el año próximo en la media siguiente. Podremos observar que en cada uno de los cálculos siempre consideraremos los datos de un año completo.

Ejercicio 5

En la siguiente tabla presentamos el número total de viajeros trasladados en los servicios de transporte público en la Comunidad Valenciana, detallados por meses, de los años 2006 al 2010.

	2006	2007	2008	2009	2010
Enero	12340	12542	12622	11869	10283
Febrero	11850	12156	12202	11365	10981
Marzo	13721	13729	10457	12024	12207
Abril	10919	11612	12960	10469	10549
Mayo	13495	13777	12713	12038	12076
Junio	13029	13094	12619	12314	11783
Julio	12118	12364	12351	11490	10621
Agosto	8803	8814	8846	8027	7698
Septiembre	12148	11768	12057	10964	10705
Octubre	13141	13266	13277	11956	11447
Noviembre	13307	12655	12474	11718	11557
Diciembre	11859	11624	11682	10894	10769

Solución

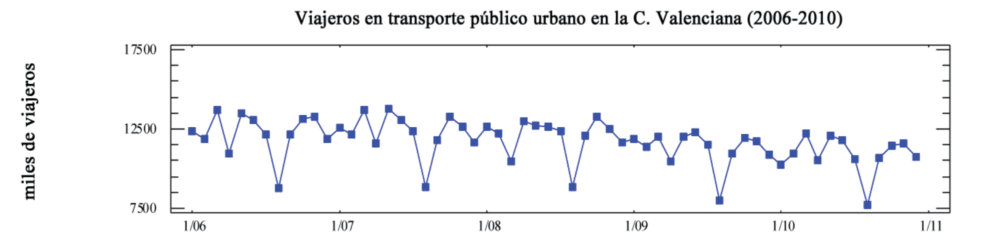
Calcula las componentes de esta serie por el método de las medias móviles. Interpreta los resultados.

Análisis de los datos para decidir el número de observaciones que deben tomarse en cada media

Para empezar la resolución por el método de las medias móviles hay que elegir el número adecuado de observaciones por incluir en cada media. Este número depende de la componente que queramos encontrar.

Comenzaremos por explicar este número en el caso de la tendencia.

Este número que denotaremos por n es el mínimo común múltiplo, MCM en adelante, el número de observaciones por año (en este caso 12 para ser observaciones mensuales) y el número de observaciones que incluye cada «período» de la gráfica (12 en este caso porque tal y como se ve, la gráfica tiene un comportamiento que se repite cada año).



Para calcular la tendencia

Así, hay que hacer las medias aritméticas de 12 datos, de tal modo que en el primer caso, calculemos su media:

$$y_{1-12} = \frac{12340+11850+13721+10919+...+12148+13141+13307+11859}{12} = 12227{,}5$$

y así vamos sustituyendo un dato (el primero) por el siguiente dato de la serie, de tal manera que siempre hacemos la media de 12 datos, es decir, determinamos por el método de las medias móviles de 12 datos la segunda columna que utilizaremos para calcular la tendencia de la serie. Se puede observar que siempre se trata de la media de los datos de todo un año, aunque esta colección va empezando en cada media por los diferentes meses:

Podemos observar como hemos calculado los siguientes términos de la segunda columna:

$$y_{2-13} = \frac{11850+13721+10919+...+13141+13307+11859+12542}{12} = 12244{,}3333$$

$$y_{3-14} = \frac{13721+10919+13495+...+13307+11859+12542+12156}{12} = 12269{,}8333$$

$$y_{4-15} = \frac{10919+13495+13029+...+11859+12542+12156+13729}{12} = 12270{,}5$$

$$y_{5-16} = \frac{113495+13029+12118+...+11859+12542+12156+13729}{12} = 12328{,}25$$

y así, sucesivamente, se calculan los valores restantes de la segunda columna que hemos denotado por $\bar{y}$.

Un tema que hay que especificar es que estas medias corresponden a 12 meses (número par), por lo que es difícil hacer corresponder estas medias calculadas a ninguno de los períodos de los que partimos, ya que no tendremos un período que corresponda al centro de las observaciones promediado. En estos casos en que el promedio corresponde a un número par de datos, hay que calcular la tercera columna donde tendremos las medias móviles centradas y, además, es la media aritmética de cada dos datos consecutivos de la segunda, columna para hacerlos corresponder a un mes en particular.

Explicamos cómo obtener los primeros valores de la tercera columna, denotada por T_{ij}, los valores de la que consideramos que explican la «tendencia» de la serie y nos permite ver el comportamiento de los valores de la serie a largo plazo. A diferencia del método del ajuste analítico, la tendencia es una secuencia numérica extraída de los valores de los datos y no disponemos de una expresión algebraica para obtener un valor de tendencia para cada año.

En este método de las medias móviles obtenemos un valor diferente que hace referencia a cada mes de cada año, exceptuando los valores iniciales y finales que no podemos calcular.

Enero 2006	12340		
Febrero 2006	11850		
Marzo 2006	13721		
Abril 2006	10919		
Mayo 2006	13495		
Junio 2006	13029	12227,5	
Julio 2006	12118	12244,3333	12235,9167
Agosto 2006	8803	12269,8333	12257,0833
Septiembre 2006	12148	12270,5	12270,1667
Octubre 2006	13141	12328,25	12299,375
Noviembre 2006	13307	12351,75	12340
Diciembre 2006	11859	12357,1667	12354,4583
Enero 2007	12542	12377,6667	12367,4167
Febrero 2007	12156	12378,5833	12378,125
Marzo 2007	13729	12346,9167	12362,75
Abril 2007	11612	12357,3333	12352,125
Mayo 2007	13777	12303	12330,1667
Junio 2007	13094	12283,4167	12293,2083
Julio 2007	12364	12290,0833	12286,75
Agosto 2007	8814	12293,9167	12292

Septiembre 2007	11768	12021,25	12157,5833
Octubre 2007	13266	12133,5833	12077,4167
Noviembre 2007	12655	12044,9167	12089,25
Diciembre 2007	11624	12005,3333	12025,125
Enero 2008	12622	12004,25	12004,7917
Febrero 2008	12202	12006,9167	12005,5833
Marzo 2008	10457	12031	12018,9583
Abril 2008	12960	12031,9167	12031,4583
Mayo 2008	12713	12016,8333	12024,375
Junio 2008	12619	12021,6667	12019,25
Julio 2008	12351	11958,9167	11990,2917
Agosto 2008	8846	11889,1667	11924,0417
Septiembre 2008	12057	12019,75	11954,4583
Octubre 2008	13277	11812,1667	11915,9583
Noviembre 2008	12474	11755,9167	11784,0417
Diciembre 2008	11682	11730,5	11743,2083
Enero 2009	11869	11658,75	11694,625
Febrero 2009	11365	11590,5	11624,625
Marzo 2009	12024	11499,4167	11544,9583
Abril 2009	10469	11389,3333	11444,375
Mayo 2009	12038	11326,3333	11357,8333
Junio 2009	12314	11260,6667	11293,5
Julio 2009	11490	11128,5	11194,5833
Agosto 2009	8027	11096,5	11112,5
Septiembre 2009	10964	11111,75	11104,125
Octubre2009	11956	11118,4167	11115,0833
Noviembre 2009	11718	11121,5833	11120
Diciembre 2009	10894	11077,3333	11099,4583
Enero 2010	10283	11004,9167	11041,125
Febrero 2010	10981	10977,5	10991,2083
Marzo 2010	12207	10955,9167	10966,7083
Abril 2010	10549	10913,5	10934,7083
Mayo 2010	12076	10900,0833	10906,7917
Junio 2010	11783	10889,6667	10894,875
Julio 2010	10621		

Agosto 2010	7698		
Septiembre 2010	10705		
Octubre 2010	11447		
Noviembre 2010	11557		
Diciembre 2010	10769		

Podemos presentar los cálculos de los primeros valores de la tendencia:

$$T_{juliol2006} = \frac{12227,5 + 12244,3333}{2} = 12235,9167$$

$$T_{agost2006} = \frac{12244,3333 + 12269,8333}{2} = 12257,0833$$

$$T_{setembre2006} = \frac{12269,8333 + 12270,5}{2} = 12270,1667$$

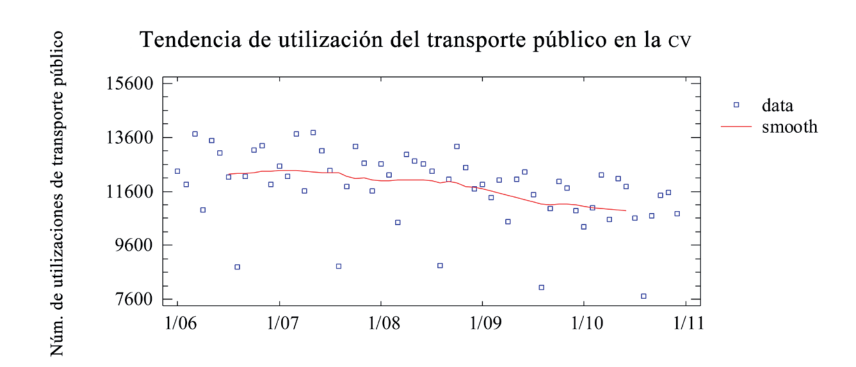
y así podemos comprobar los valores del resto de la columna.

Habría que explicar mejor que los valores de la segunda y tercera columnas, quizás quedarían mejor referidos al período, si las presentaremos como lo hacemos en la tabla siguiente. Hemos calculado los promedios de 12 valores y los hacemos corresponder en el «centro» de este período, que situamos en la celda de la tabla que hemos previsto intercalando una fila entre cada dos filas de datos originales, y que después permitirán ver más claramente los datos de este procedimiento de medias móviles centradas. No hemos presentado los cálculos de toda la serie por motivos de espacio, como es evidente. Sin embargo, sí queremos mostrarlos con los primeros valores de la serie:

	y_k		T_k
Enero 2006	12340		
Febrero 2006	11850		
Marzo 2006	13721		
Abril 2006	10919		
Mayo 2006	13495		
Junio 2006	13029		
		12227,5	
Julio 2006	12118		12235,9167
		12244,3333	
Agosto 2006	8803		12257,0833

		12269,8333	
Septiembre 2006	12148		12270,1667
		12270,5	
Octubre 2006	13141		12299,375
		12328,25	
Noviembre 2006	13307		12340
		12351,75	
Diciembre 2006	11859		12354,4583
		12357,1667	
Enero 2007	12542		12367,4167
		12377,6667	
Febrero 2007	12156		12378,125
		12378,5833	
Marzo 2007	13729		12362,75
		12346,9167	
Abril 2007	11612		12352,125
		12357,3333	
Mayo 2007	13777		12330,1667
		12303	
Junio 2007	13094		12293,2083
		12283,4167	
Julio 2007	12364		12286,75
		12290,0833	
Agosto 2007	8814		12292
		12293,9167	
Septiembre 2007	11768		12157,5833
		12021,25	
Octubre 2007	13266		12077,4167
		12133,5833	
Noviembre 2007	12655		12089,25
		12044,9167	
Diciembre 2007	11624		12025,125

Se puede observar que los primeros y últimos datos no tienen sus medias correspondientes por la técnica utilizada. Podremos interpretar mejor el sentido de estos datos calculados con la representación de los valores encima de la serie, que nos permite interpretar el comportamiento a largo plazo de los valores de la utilización del transporte público en la Comunidad Valenciana:



Podemos ver que a largo plazo, si consideramos los datos que analizamos, la utilización del transporte público en la Comunidad Valenciana parece disminuir. Tiene un comportamiento decreciente si observamos en el gráfico anterior la línea roja que representa los valores de la tendencia.

Para calcular la componente estacional

Para continuar el análisis debemos determinar ahora la componente estacional. A tal fin, es necesario que calculamos por un lado los valores de las medias aritméticas de los valores de cada mes. Con este propósito, situaremos los datos originales distribuidos en filas por períodos y en columnas por años, como se puede ver en la tabla siguiente. En la columna de la derecha hemos calculado las medias ya citadas,

$$\bar{y}_k = \frac{\sum_i y_{ij}}{N}.$$

	2006	2007	2008	2009	2010	$\bar{y}_k$
Enero	12340	12542	12622	11869	10283	11931,2
Febrero	11850	12156	12202	11365	10981	11710,8
Marzo	13721	13729	10457	12024	12207	12427,6
Abril	10919	11612	12960	10469	10549	11301,8
Mayo	13495	13777	12713	12038	12076	12819,8
Junio	13029	13094	12619	12314	11783	12567,8
Julio	12118	12364	12351	11490	10621	11788,8
Agosto	8803	8814	8846	8027	7698	8437,6
Septiembre	12148	11768	12057	10964	10705	11528,4
Octubre	13141	13266	13277	11956	11447	12617,4
Noviembre	13307	12655	12474	11718	11557	12342,2
Diciembre	11859	11624	11682	10894	10769	11365,6

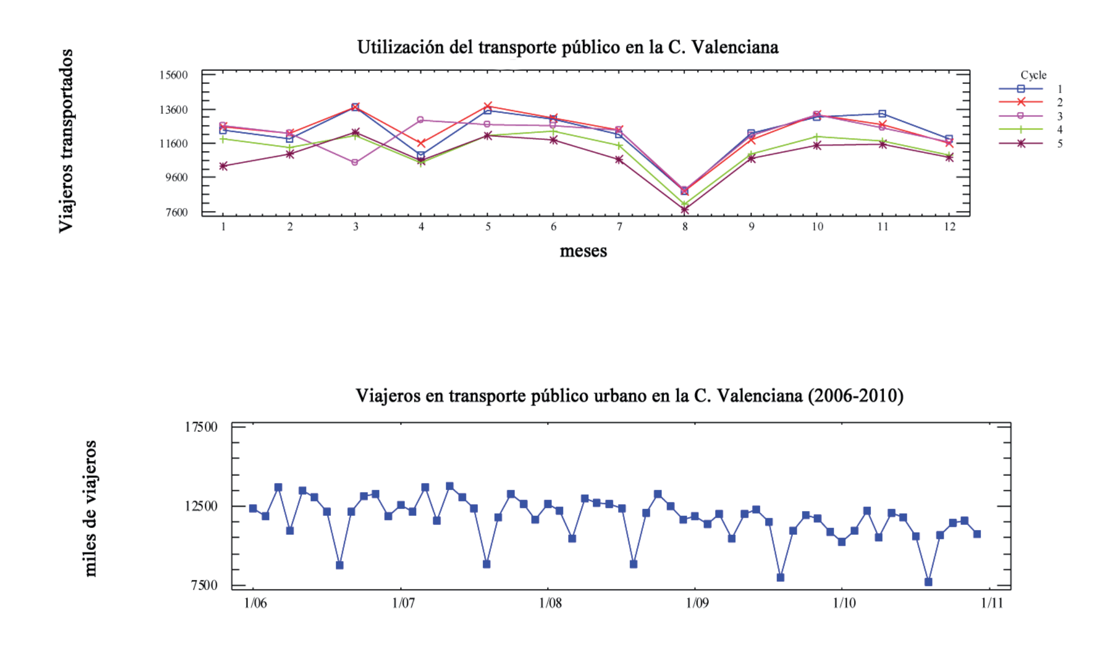
Por otra parte, vamos a ubicar en la tabla siguiente los valores de las medias móviles que hay que calcular con un número de datos p, que es el número de datos que tenemos por año. En este problema, podremos aprovechar los valores de la tercera columna del cálculo de la tendencia, ya que en este caso también utilizamos medias móviles con $p = 12$. Así pues, redistribuiremos los valores de dichos medios en una tabla donde cada celda indicará el valor que corresponde a cada mes y año.

	2006	2007	2008	2009	2010	E_k
Enero		12367,4167	12004,7917	11694,625	11041,125	11776,9896
Febrero		12378,125	12005,5833	11624,625	10991,2083	11749,8854
Marzo		12362,75	12018,9583	11544,9583	10966,7083	11723,3437
Abril		12352,125	12031,4583	11444,375	10934,7083	11690,6667
Mayo		12330,1667	12024,375	11357,8333	10906,7917	11654,7917
Junio		12293,2083	12019,25	11293,5	10894,875	11625,2083
Julio	12235,9167	12286,75	11990,2917	11194,5833		11926,8854
Agosto	12257,0833	12292	11924,0417	11112,5		11813,8568
Septiembre	12270,1667	12157,5833	11954,4583	11104,125		11757,5058
Octubre	12299,375	12077,4167	11915,9583	11115,0833		11716,491
Noviembre	12340	12089,25	11784,0417	11120		11677,4457
Diciembre	12354,4583	12025,125	11743,2083	11099,4583		11636,3093

Para calcular la componente estacional habrá que restar las columnas finales de estas dos tablas, $e_k = \bar{y}_k - E_k$, lo que nos permitirá explicar cuáles son los meses de mayor y menor utilización del transporte público en la Comunidad Valenciana.

	e_k
Enero	154,2104
Febrero	–39,0854
Marzo	704,256275
Abril	–388,86665
Mayo	1165,00833
Junio	942,591675
Julio	–138,085425
Agosto	–3376,25678
Septiembre	–229,105845
Octubre	900,908964
Noviembre	664,754316
Diciembre	–270,709321

Así pues, veamos los resultados en la siguiente columna, donde podemos destacar que el mes de mayor utilización del transporte público en la Comunidad Valenciana se produce en mayo, seguido también de los meses de junio y octubre, por lo contrario, se ve que el mes que destaca porque su valor es muy inferior a la media es el de agosto. Esta interpretación podemos comprobarla porque los máximos y mínimos de la gráfica de la serie coinciden con estos valores de la componente estacional.



Para calcular la componente residual o errática

Para calcular la componente errática o residual hay que restar a cada valor inicial el valor de la tendencia correspondiente a cada celda en los casos en que esta existe (no a los primeros y últimos valores de la serie), y también le restamos el valor de la componente estacional que corresponde a cada período.

En la siguiente tabla se indican estos cálculos. Así, la componente errática para cada período será:

	2006	2007	2008	2009	2010
Enero		12542–12367,4167––154,2104	12622–12004,7917 –154,2104	11869–11694,625 –154,2104	10283 –11041,125 –154,210
Febrero		12156–12378,125 +39,0854	12202–12005,5833 +39,0854	11365–11624,625 +39,0854	10981–10991,2083 +39,0854
Marzo		13729–12362,75 –704,256275	10457–12018,9583 –704,256275	12024–11544,9583 –704,256275	12207–10966,7083 –704,256275
Abril		11612–12352,125 +388,86665	12960–12018,9583 +388,86665	10469–11444,375 +388,86665	10549–10934,7083 +388,86665
Mayo		13777–12330,1667 –1165,00833	12713–12024,375 –1165,00833	12038–11357,8333 –1165,00833	12076–10906,7917 –1165,00833
Junio		13094–12293,2083 –942,591675	12619–12019,25 –942,591675	12314–11293,5 –942,591675	11783–10894,875 –942,591675
Julio	12118 – 12235,9167 +138,085425	12364–12286,75 +138,085425	12351–11990,2917 +138,085425	11490–11194,5833 +138,085425	
Agosto	8803 – 12257,0833 +3376,25678	8814–12292 +3376,25678	8846–11924,0417 +3376,25678	8027 – 11112,5 +3376,25678	
Septiembre	12148 – 12270,1667 +229,105845	11768–12157,5833 +229,105845	12057–11954,4583 +229,105845	10964–11104,125 +229,105845	
Octubre	13141–12299,375 –900,908964	13266–12077,4167 –900,908964	13277–11915,9583 –900,908964	11956–11115,0833 –900,908964	
Noviembre	13307 – 12340 – 664,754316	12655–12089,25 –664,754316	12474–11784,0417 -664,754316	11718–11120 –664,754316	
Diciembre	11859–12354,4583 +270,709321	11624–12025,125 +270,709321	11682–11743,2083 +270,709321	10894–11099,4583 +270,709321	

Así pues, podemos concluir que la variable residual o errática correspondiente a cada período son los valores de la siguiente tabla:

	2006	2007	2008	2009	2010
Enero		20,3729	462,9979	20,1646	–912,335
Febrero		–183,0396	235,5021	–220,5396	28,8771
Marzo		661,993725	–2266,21458	–225,214575	536,035425
Abril		–351,25835	1329,90835	–586,50835	3,15835
Mayo		281,82497	–476,38333	–484,84163	4,19997
Junio		–141,799975	–342,841675	77,908325	–54,466675
Julio	20,168725	215,335425	498,793725	433,502125	
Agosto	–77,82652	–101,74322	298,21508	290,75678	
Septiembre	106,939145	–160,477455	331,647545	88,980845	
Octubre	–59,283964	287,674336	460,132736	–59,992264	
Noviembre	302,245684	–99,004316	25,203984	–66,754316	
Diciembre	–224,748979	–130,415679	209,501021	65,251021	

Para facilitar la interpretación de los datos de esta tabla, hemos resaltado en rojo los valores que corresponden a períodos donde la utilización del transporte público ha sido por encima de lo esperado atendiendo al modelo que hemos encontrado, y los valores reseñados en azul corresponden a períodos con valores que están por debajo de las previsiones, siempre según el modelo que hemos estudiado.

Podemos destacar que los valores que se encuentran de alguna manera «fuera del modelo» corresponden, en todos los años de nuestra tabla, a los meses de marzo y abril. Podríamos explicar que, tal vez, hay que considerar que en estos meses se producen fiestas importantes en la Comunidad y que las vacaciones de Semana Santa y Pascua son también en estos meses de manera oscilante y no en un mes concreto, lo que hubiera facilitado el estudio de la componente estacional.

Hacemos esta consideración porque hemos observado que la componente estacional nos permite interpretar que los meses que corresponden a vacaciones escolares y/o laborales (julio, agosto, septiembre y diciembre) tienen valores que indican menos utilización del transporte público. También el mes de abril está en este grupo y puede variar, según el año, el comportamiento de marzo y abril como se puede interpretar de los datos de la componente errática o residual.

Ejercicio 6

Realiza la gráfica de la siguiente serie que indica los miles de kilos de fruta comercializada por trimestres en los últimos 4 años.

	2008	2009	2010	2011
1.er trimestre	10	23	12	29
2.o trimestre	11	27	11	28
3.er trimestre	9	25	10	21
4.er trimestre	8	20	8	23

Observa las siguientes tablas que presenten los cálculos que hemos hecho para obtener la tendencia y la componente estacional de dicha serie.

datos		
10		
11		
9		
8		
23		
27		
	16,875	
25		16,9375
	17	
20		17
	17	
12		17,375
11		
	17,875	

datos		
10		
11		
9		
	12,75	
8		
	16,75	
23		18,75
27		
	23,75	
25		22,375
	21	
20		19
	17	
12		15,125
	13,25	
11		11,75
	10,25	

10		17,625
	17,375	
8		
29		
28		
21		
23		

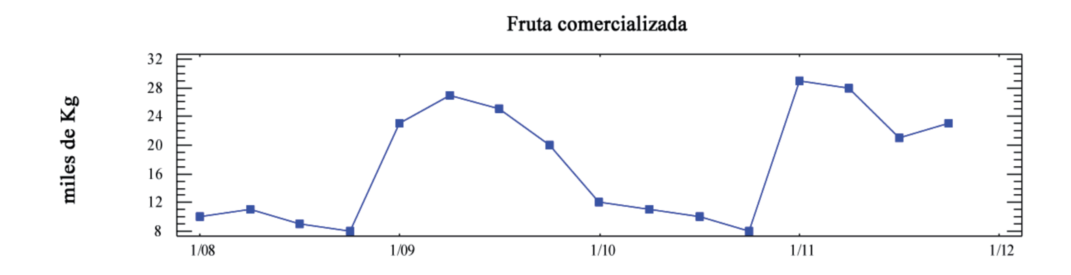
10		
8		16,625
	18,75	
29		20,125
28		
21		
23		

Solución

a) Identifica el método empleado, añade aquellos datos que faltan en las tablas, comentando el procedimiento de cálculo que hay que hacer en ellas, justificándolos.

Gráfico y visualización del modelo

Para empezar, presentaremos el gráfico de los datos de la serie original:



y podemos ver que la periodicidad del fenómeno, hace que los datos repitan el «patrón» cada dos años.

Cálculo de la tendencia

Nos piden que identifiquemos el procedimiento para obtener las componentes de la serie. Parece bastante evidente que estamos trabajando por el método de las medias móviles y habrá que ver cuál es el número de datos que hay que tomar para hacer cada media en el cálculo de la tendencia.

Para decidirnos hay que considerar dos números: el número de datos de cada «período» de la gráfica antes presentada (en este caso, es evidente, que es 8) y el número de observaciones por año (en este caso, es 4 porque son datos trimestrales).

Por lo que el número de datos por coger es el MCM (4, 8) = 8. Pero esta decisión permite ver que la tabla de la izquierda está construida para calcular la tendencia, haciendo en la columna primera las medias con $p = 8$, y a la derecha el posterior centrado que hay que hacer por tratarse de un número par de datos.

También observamos que está incompleta. Presentamos la tabla finalizada y a continuación los cálculos que hemos hecho para completarla.

La primera media que falta en la primera columna la hemos obtenido tomando los primeros 8 datos (lo indicaremos en el subíndice). Así:

$$y_{1-8} = \frac{10+11+9+8+23+27+25+20}{8} = 16{,}625$$

del mismo modo, continuaremos rellenando las siguiente casillas de la misma columna. Indicamos al subíndice el ordinal de los datos extremos del intervalo de datos que tomaremos en cada media. Hay que recordar que estamos empleando un método de «suavizado» para medias móviles.

$$y_{2-9} = \frac{11+9+8+23+27+25+20+12}{8} = 16{,}875$$

$$y_{3-10} = \frac{9+8+23+27+25+20+12+11}{8} = 16{,}875 \text{ (que ya figura en la tabla)}$$

y así podemos ir comprobando las celdas las ya calculadas hasta que llegamos a la media siguiente que está vacía:

$$y_{6-13} = \frac{27+25+20+12+11+10+8+29}{8} = 17{,}75$$

y también falta la última celda que corresponde a la media siguiente (que calculamos con los datos de los últimos dos años):

$$y_{9-16} = \frac{12+11+10+8+29+28+21+23}{8} = 17{,}75$$

Como podemos ver en la tabla anterior, esta columna no se relaciona todavía con los valores de la tendencia, ya que las medias obtenidas no corresponden a los períodos (trimestres) de los que tenemos los datos. Como se trata de medias obtenidas con un número par de datos se hará un posterior centrado para evitar este problema. A tal fin, hay que hacer la media aritmética de cada dos datos (medias de la columna primera).

Para completar esta primera columna, que ya corresponde a los valores de la tendencia para cada período (T_{ik}), hemos empezado llenando las primeras casillas que faltaban a partir de los primeros datos de la columna anterior. Así, tenemos:

$$T_{09/1} = \frac{16,625 + 16,875}{2} = 16,75$$

$$T_{09/2} = \frac{16,875 + 16,875}{2} = 16,875$$

los siguientes datos de la tabla están ya calculados hasta llegar a la siguiente celda la vacía que corresponde a:

$$T_{10/2} = \frac{17,75 + 17,875}{2} = 17,8125$$

y la última, obtenida a partir de las últimas medias:

$$T_{10/4} = \frac{17,375 + 17,75}{2} = 17,5625$$

Podemos ver que el cálculo de la tendencia por este método impide obtener la tendencia de los primeros y últimos períodos por carencia de datos para calcular las medias móviles y el posterior centrado.

b) Calcula la componente estacional

El cálculo de la componente estacional por el método de las medias móviles está reflejado en la tabla de la derecha. En este caso el número de observaciones que nos hace falta coger para cada media, es el número de observaciones que tenemos por año. En este caso los datos son trimestrales, pues, hay que coger 4 datos en cada media.

Indicamos a continuación los cálculos de las medias que faltan en las celdas vacías e indicaremos el subíndice con el ordinal de los datos del intervalo de valores que tomamos en cada una.

Así, empezaremos por las primeras celdas que podemos calcular en la primera columna. Los resultados están en rojo en la tabla siguiente:

$$\overline{y}_{1-4} = \frac{10 + 11 + 9 + 8}{4} = 9,5$$

$$\overline{y}_{2-5} = \frac{11 + 9 + 8 + 23}{4} = 12,75 \text{ (ya está en la tabla)}$$

Podemos comprobar los siguientes resultados hasta llegar a las celdas:

$$\overline{y}_{4-7} = \frac{8 + 23 + 27 + 25}{4} = 20,75$$

$$\bar{y}_{10-13} = \frac{11+10+8+29}{4} = 14,5$$

Y las de las últimas columnas:

$$\bar{y}_{12-15} = \frac{8+29+28+21}{4} = 21,5$$

$$\bar{y}_{13-16} = \frac{29+28+21+23}{4} = 25,25$$

Como también se trata de un número par de datos ($p = 4$) habrá también que hacer el centrado, al igual que en la tabla de la izquierda. Para llenar la segunda columna recordemos que deberemos hacer la media aritmética de cada dos medias. Indicaremos el subíndice del año y trimestre al que hacer corresponder cada resultado. Así, la primera celda, denotada por $\bar{y}'_{08/3}$ corresponde a la media móvil del tercer trimestre del año 2008.

$$\bar{y}'_{08/3} = \frac{9,5+12,75}{2} = 11,125$$

$$\bar{y}'_{08/4} = \frac{12,75+16,75}{2} = 14,75$$

$$\bar{y}'_{09/2} = \frac{20,75+23,75}{2} = 22,25$$

$$\bar{y}'_{10/3} = \frac{10,25+14,5}{2} = 12,375$$

$$\bar{y}'_{11/2} = \frac{21,5+25,25}{2} = 23,375$$

Todos estos resultados se pueden ver en la tabla siguiente:

datos		T_k
10		
11		
9		
8		
	16,625	

datos		$\bar{y}'_k$
10		
11		
	9,5	
9		**11,125**
	12,75	
8		**14,75**
	16,75	

23		**16,75**
	16,875	
27		**16,875**
	16,875	
25		16,9375
	17	
20		17
	17	
12		17,375
	17,75	
11		**17,8125**
	17,875	
10		17,625
	17,375	
8		**17,5625**
	17,75	
29		
28		
21		
23		

23		18,75
	20,75	
27		**22,25**
	23,75	
25		22,375
	21	
20		19
	17	
12		15,125
	13,25	
11		11,75
	10,25	
10		**12,375**
	14,5	
8		16,625
	18,75	
29		20,125
	21,5	
28		**23,375**
	25,25	
21		
23		

Pero, recordemos que estas medias de la columna de la derecha de la tabla derecha no son más que uno de los pasos del cálculo de la componente estacional:

Hay que hacer, en parte, el cálculo de las medias de los datos originales por períodos. Las denotaremos por $\overline{y}_k$ y están en la columna derecha de la siguiente tabla:

	2008	2009	2010	2011	$\overline{y}_k$
1.er trimestre	10	23	12	29	18,5
2.º trimestre	11	27	11	28	19,25
3.er trimestre	9	25	10	21	16,25
4.º trimestre	8	20	8	23	14,75

Por otra parte, situaremos las medias móviles de la tabla de la derecha, ya cumplimentada, en esta distribución por años y períodos, haciendo corresponder cada valor al período adecuado. Veamos:

	2008	2009	2010	2011	$\overline{y}_k$	$\overline{E}_k$	e_k
1.er trimestre		18,75	15,125	20,125	18,5	18	0,5
2.o trimestre		22,25	11,75	23,375	19,25	19,125	0,125
3.er trimestre	11,125	22,375	12,375		16,25	15,292	0,958
4.o trimestre	14,75	19	16,625		14,75	16,792	-2,042

Mantenemos la columna de la tabla anterior con los valores de $\overline{y}_k$, y en la columna siguiente también calculamos la media aritmética de estos valores por filas o períodos, considerando el número de valores que disponemos. Hay que ver que quedan celdas vacías. Los denotaremos por $\overline{E}_k$.

Anotemos luego cómo calcularlas:

$$\overline{E}_1 = \frac{18{,}75 + 15{,}125 + 20{,}125}{3} = 18$$

$$\overline{E}_2 = \frac{22{,}25 + 11{,}75 + 23{,}375}{3} = 19{,}125$$

$$\overline{E}_3 = \frac{11{,}125 + 22{,}375 + 12{,}375}{3} = 15{,}292$$

$$\overline{E}_4 = \frac{14{,}75 + 19 + 16{,}625}{3} = 16{,}792$$

y en la última columna ya podemos obtener la componente estacional e_k, restando estas columnas que acabamos de explicar: $e_k = \overline{y}_k - \overline{E}_i$.

$$e_1 = 18{,}5 - 18 = 0{,}5$$

$$e_2 = 19{,}25 - 19{,}125 = 0{,}125$$

$$e_3 = 16{,}25 - 15{,}292 = 0{,}958$$

$$e_4 = 14{,}75 - 16{,}792 = -2{,}042$$

Podemos interpretar, si analizamos estos resultados, que los datos del cuarto trimestre son muy inferiores a los del resto y destacaríamos los valores del tercer trimestre por encima de la media global.

c) Calcula la componente errática o residual.

Para calcular la componente errática o residual, considerando que trabajamos con un modelo aditivo, restaremos a cada dato original los valores de la tendencia y la componente estacional. Así, distribuiremos también la tabla de los valores de la tendencia por períodos y años. En la siguiente tabla tenemos esta distribución, por lo que se puede ver que solo podremos calcular la componente residual de las dos columnas centrales.

	2008	2009	2010	2011
1.er trimestre		16,75	17,375	
2.o trimestre		16,875	17,8125	
3.er trimestre		16,9375	17,625	
4.o trimestre		17	17,5625	

Así, considerando que $r_k = y_k - T_k - e_k$, tenemos los siguientes resultados:

	2008	2009	2010	2011
1.er trimestre		23 – 16,75 – 0,5 = 5,75	12 – 17,375 – 0,5 = –5,875	
2.o trimestre		27 – 16,875 – 0,125 = 10	11 – 17,8125 – 0,125 = –6,9375	
3.er trimestre		25 – 16,9375 – 0,958 = 7,1045	10 – 17,625 – 0,958 = –8,583	
4.o trimestre		20 – 17 + 2,042 = 5,042	8 – 17,5625 + 2,042 = –7,5205	

En los valores que obtenemos en esta tabla, destaca mucho el comportamiento de los datos con fluctuaciones muy importantes de cada año, como podíamos ver en la gráfica inicial. Podemos ver que un año tiene los valores muy bajos y el siguiente mucho mayores. Esta circunstancia es la que podemos ver reflejada en la tabla anterior de la componente residual, que en esta serie, lo que se destaca es más bien ese comportamiento cíclico bianual.

Ejercicio 7

La siguiente serie cronológica muestra el número de nuevas contrataciones de cierta superficie comercial por cuatrimestres en los años que se indican:

	2007	2008	2009	2010	2011
1.er cuatrimestre	41	39	35	21	22
2.o cuatrimestre	37	33	27	15	16
3.er cuatrimestre	36	30	16	12	13

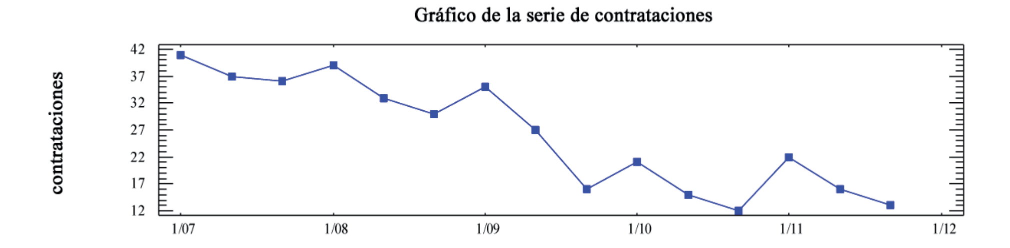
a) Realiza la gráfica de la serie y, suponiendo un modelo aditivo, calcula por el método del ajuste analítico las componentes de esta serie, e interpreta cada uno de los resultados obtenidos.
b) Haz las previsiones que podemos esperar para los años 2012 y 2013.
c) Realiza el análisis de la serie por el método de las medias móviles, estimando también las componentes de la serie.
d) Compara los resultados del análisis por los dos métodos empleados.

Solución

a) Realiza la gráfica de la serie y, suponiendo un modelo aditivo, calcula por el método del ajuste analítico las componentes de esta serie, e interpreta cada uno de los resultados obtenidos.

Identificación del modelo y gráfica

Siempre hay que empezar con una gráfica de los datos para ver el patrón de comportamiento de la serie y nos confirme que podemos aplicarle un ajuste analítico de tipo aditivo:



Viendo esta gráfica solo cabe esperar una tendencia decreciente y también una componente estacional bastante marcada por la «periodicidad», que podemos ver en la gráfica (eje X) que coincide con el intervalo anual que corresponde cada tres datos.

Cálculo de la tendencia

Para calcular la componente de la tendencia llenaremos las casillas de la parte inferior de la tabla siguiente para ajustar una recta de regresión en las medias cuatrimestrales anuales que figuran en la fila $\bar{y}_i = \frac{\sum_k y_{ik}}{m}$.

Hacer notar que $\bar{y}_i$ es la media de los valores de cada columna. Recordemos tan solo que cada dato de una serie temporal la representamos por dos subíndices y_{ik}, donde i hace referencia al año y k hace referencia al período. Además, m indica el número de filas de la tabla o número de períodos de los que tenemos datos para cada año. En nuestro caso, $m = 3$ porque un año tiene 3 cuatrimestres.

Podemos ver que en la fila siguiente hay una escalera i que hace referencia a los años para simplificar los cálculos; es recomendable poner el valor 0 en la columna central porque simplificará mucho el cálculo del sistema que hay que plantear para obtener los coeficientes de la recta de tendencia.

A continuación llenaremos las filas tercera y cuarta que hacen referencia a $\overline{y}_i \cdot i$ y i^2, considerando los valores de las filas anteriores.

Después sumaremos las filas y obtendremos los valores que podemos encontrar en la columna «totales».

	2007	2008	2009	2010	2011	
1.er cuatrimestre	41	39	35	21	22	
2.º cuatrimestre	37	33	27	15	16	
3.er cuatrimestre	36	30	16	12	13	TOTALES
$\overline{y}_i$	38	34	26	16	17	131
i	–2	–1	0	1	2	0
$\overline{y}_i \cdot i$	–76	–34	0	16	34	–60
i^2	4	1	0	1	4	10

Utilizaremos estos valores «totales» para resolver el sistema que se plantea a continuación:

$$\begin{cases} \sum_i \overline{y}_i = Na + b\sum_i i \\ \sum_i \overline{y}_i \cdot i = a\sum_i i + b\sum_i i^2 \end{cases}$$

donde N es el número de años de los que tenemos datos en la tabla (en este caso, 4 años que corresponden a las 4 columnas) y los coeficientes a y b son los coeficientes de la recta de regresión que denominamos recta de tendencia, la formula es $T_i = a + ib$, donde i hace referencia al año que se indica en la escala de la tabla superior.

$$\begin{cases} \sum_i \overline{y}_i = Na + b\sum_i i \\ \sum_i \overline{y}_i \cdot i = a\sum_i i + b\sum_i i^2 \end{cases} \Rightarrow \begin{cases} 131 = 5a + 0b \\ -60 = 0a + 10b \end{cases}$$

Resolveremos el sistema aislando en cada ecuación el valor de los coeficientes *a* i *b*. La tarea se facilita por la situación del 0 de la escala en la columna central en cuanto su número es impar.

Los valores que hemos calculado son: $a = \frac{131}{5} = 26,2$ y $b = \frac{-60}{10} = -6$, con los que podemos concluir que $Ti = 26,2 - 6_i$.

Para interpretar esta componente de la serie, T_i, que se llama tendencia y que nos explica el comportamiento del fenómeno a «largo plazo», nos fijamos en el valor de la pendiente $b = -6$, que por ser negativa nos permite explicar que las contrataciones van decreciendo año tras año, y podemos además detallar que el valor de la media de gastos cuatrimestrales del departamento, $\overline{y}_i$, ha disminuido en 6 cada año.

Cálculo de la componente estacional

La segunda componente de la serie que hay que calcular es la componente estacional, que denotaremos por e_k, que nos permitirá analizar el comportamiento del fenómeno por períodos dentro del año (en nuestro ejercicio, por cuatrimestres), valorando en cuál de ellos los valores están por encima o por debajo de un valor global que denotaremos por *M* y que llamaremos media global corregida.

Los valores de e_k estarán expresados en valores absolutos y en las mismas unidades que los datos de la tabla original (en este ejercicio en miles de euros).

Para abordar los cálculos de e_k, trabajaremos ampliando la tabla en diferentes columnas hacia la derecha de la tabla original, ya que nos interesa hacer un trabajo por períodos, en este caso, por cuatrimestres.

En la columna primera calcularemos la media de los valores de cada cuatrimestre y esta media la denotaremos por $\overline{y}_k = \frac{\sum_i y_{ij}}{N}$.

	2007	2008	2009	2010	2011		$\overline{y}_k$	$\overline{y}_k'$	e_k	I_k
1.er cuatrimestre	41	39	35	21	22		31,6	31,6	3,4	112,06 %
2.o cuatrimestre	37	33	27	15	16		25,6	27,6	–0,6	97,87 %
3.er cuatrimestre	36	30	16	12	13	TOTALES ↓	21,4	25,4	–2,8	90,07 %
								$M = 28,2$		

En la siguiente columna calcularemos las medias corregidas $\overline{y}_k^{'}$ (donde eliminamos en cada dato el valor proporcional a la tendencia que le podemos asignar a cada período), suponiendo que este decrecimiento $b = -6$ ha sido constante a lo largo de los períodos del año.

En este ejercicio, $b/m = -6/3 = -2$. Al ser una tendencia negativa, los valores de las medias corregidas $\overline{y}_k^{'}$ serán mayores que las medias originales $\overline{y}_k$, ya que estas se calculan así:

$$\overline{y}_k^{'} = \overline{y}_k - \frac{b}{m}(k-1)$$

Así:

$$\overline{y}_1^{'} = \overline{y}_1 - \frac{b}{m}(1-1) = 31{,}6 - (-2)(1-1) = 31{,}6$$

$$\overline{y}_2^{'} = \overline{y}_2 - \frac{b}{m}(2-1) = 25{,}6 - (-2)(2-1) = 27{,}6$$

$$\overline{y}_3^{'} = \overline{y}_3 - \frac{b}{m}(3-1) = 21{,}4 - (-2)(3-1) = 25{,}4$$

A continuación, en una celda la inferior, calculamos M, la media de las medias corregidas, ya que:

$$M = \frac{\sum_k \overline{y}_k^{'}}{m} = \frac{31{,}6 + 27{,}6 + 25{,}4}{3} = 28{,}2$$

y representa el valor de referencia para analizar los valores de los períodos, mediante $e_k = \overline{y}_k^{'} - M$, que calcularemos en la siguiente columna.

$$e_1 = \overline{y}_1^{'} - M = 31{,}6 - 28{,}2 = 3{,}4$$

$$e_2 = \overline{y}_2^{'} - M = 27{,}6 - 28{,}2 = -0{,}6$$

$$e_3 = \overline{y}_3^{'} - M = 25{,}4 - 28{,}2 = -2{,}8$$

La componente estacional nos explicita el comportamiento por períodos en valores absolutos, indicándonos signo y cantidad en las mismas unidades que los datos originales (miles de euros). Así, diremos que en el primer cuatrimestre las contrataciones de nuestra serie tiene unos valores por encima de la media, mientras que las contrataciones del segundo cuatrimestre son aproximadamente el valor de la media y el tercer cuatrimestre son ligeramente inferiores a dicha media. El valor de la media de referencia sería $M = 28{,}2$ que podría representar una media de contrataciones cuatrimestrales global.

Otra interpretación de estos datos se puede dar con los índices estacionales $I_k = \frac{\overline{y}'_k}{M} \cdot 100$ que calculamos en la siguiente columna donde se indica en forma de porcentaje (valor relativo, donde *M* representa el 100 %) la misma información que la componente estacional, pero que al tener carácter porcentual es más fácil de presentar sin particularizar y dar el valor de *M*.

$$I_1 = \frac{\overline{y}'_1}{M} \cdot 100 = \frac{31,6}{28,2} \cdot 100 = 112,05\ \%$$

$$I_2 = \frac{\overline{y}'_2}{M} \cdot 100 = \frac{27,6}{28,2} \cdot 100 = 98,87\ \%$$

$$I_3 = \frac{\overline{y}'_3}{M} \cdot 100 = \frac{25,4}{28,2} \cdot 100 = 90,07\ \%$$

Esta componente estacional expresada en términos de porcentaje, los índices estacionales, nos permite analizar e interpretar el comportamiento de los datos de la serie, es decir, los valores de las contrataciones por cuatrimestres.

Podremos afirmar que los gastos eran mayores en el primer cuatrimestre, con valores de un 12,05 % superiores a la media anual, mientras que los valores del tercer cuatrimestre solo llegan a tener valores inferiores en un 10% aproximadamente inferiores. Cabe destacar que las contrataciones del segundo cuatrimestre tienen valores que están en torno al valor de dicha media anual global, que podríamos considerar el valor de $M = 28,2$ contrataciones.

Cálculo de la componente residual o errática

En tercer lugar, hay que calcular la componente errática o residual, que denotamos por r_{ik}, que nos permite destacar el comportamiento de algún dato y_{ik}, el valor del cual no se pueda explicar por las anteriores componentes, lo que permitirá inferir que por cualquier causa por identificar (motivos extraordinarios) este valor no está dentro del patrón de comportamiento que hemos encontrado y con el que hemos interpretado los datos originales para explicar el fenómeno. Los valores de esta componente errática deben ser pequeños en valor, variados en signo y sin regularidad ni patrón. Nos permiten ver que los datos reales no se ajustan completamente al patrón que hemos encontrado con la tendencia y la componente estacional. Es por eso que si algún valor es muy alto o bajo dejará identificar algún dato que corresponde al período de un año, el valor del cual se aleja mucho de los valores que cabría esperar.

La calcularemos así $r_{ik} = y_{ik} - T_i - e_k$ en cada celda la de la tabla. Para tal fin, primeramente hay que determinar la tendencia para cada año de la tabla $T_i = 26,2 - 6i$.

$$T_{2007} = T_{-2} = 26,2 - 6 \cdot (-2) = 38,2$$
$$T_{2008} = T_{-1} = 26,2 - 6(-1) = 32,2$$
$$T_{2009} = T_0 = 26,2 - 6 \cdot 0 = 26,2$$
$$T_{2010} = T_1 = 26,2 - 6 \cdot 1 = 20,2$$
$$T_{2011} = T_2 = 26,2 - 6 \cdot 2 = 14,2$$

Los valores de la componente estacional están en la columna de la tabla e_k. Así:

$e_1 = 3{,}4$ $\qquad$ $e_2 = -0{,}6$ $\qquad$ $e_3 = -2{,}8$

Los cálculos para cada celda están en la siguiente tabla:

	2007	2008	2009	2010	2011
1.er cuatrimestre	41 – 38,2 – 3,4 = –06	39 – 32,2– 3,4 = 3,4	35 – 26,2 – 3,4 = 5,4	21 – 20,2 – 3,4 = –2,6	22 – 14,2 – 3,4 = 4,4
2.º cuatrimestre	37 – 38,2 + 0,6 = –0,6	33 – 32,2 + 0,6 = 1,4	27– 26,2 + 0,6 = 1,4	15 – 20,2 + 0,6 = –4,6	16 – 14,2 + 0,6 = 2,4
3.er cuatrimestre	36 – 38,2 + 2,8 = 0,6	30 – 32,2 + 2,8 = 0,6	16 – 26,2 + 2,8 = –7,4	12 – 20,2 + 2,8 = –5,4	13 – 14,2 + 2,8 = 1,6

De estos resultados, podemos observar que los valores que nos llaman la atención los marcamos con color rojo. Podremos destacar los correspondientes al primer cuatrimestre de 2009, que es superior a lo que cabría esperar con $r_{2009,1} = 5{,}4$ y el tercer cuatrimestre del mismo año con un valor aún muy inferior con $r_{2011,3} = -7{,}4$. Y también el valor del tercer cuatrimestre de 2010 $r_{2010,3} = -5{,}4$ que nos indica que ese cuatrimestre es un valor inferior al que cabría esperar. Habría que estudiar si alguna circunstancia extraordinaria justifica estos valores.

b) Haz las previsiones que podemos esperar para los años 2012 y 2013.

Para hacer las estimaciones que nos piden en este apartado, recordemos que no podemos prever la componente residual, por lo que calcularemos:

$$y_{ik} = T_i + e_k$$

Considerando la escala que hemos adoptado en la tabla para los años, con el propósito de simplificar los cálculos de la tendencia, podemos establecer que si $2009 \Rightarrow i = 0$, esto comporta $2012 \Rightarrow i = 3$, lo que nos permitirá obtener el valor de la tendencia para este año: $T_{2012} = T_3 = 26,2 - 6 \cdot 3 = 8,2$ y como conocemos la componente estacional $e_1 = 3{,}4$, $e_2 = -0{,}6$ y $e_3 = -2{,}8$, podremos estimar los valores de los gastos del año 2012 por cuatrimestres:

$y_{2012,1qua} = 8{,}2 + 3{,}4 = 1{,}6$ contrataciones

$y_{2012,2qua} = 8{,}2 – 0{,}6 = 7{,}6$ contrataciones

$y_{2012,3qua} = 8{,}2 – 2{,}8 = 5{,}4$ contrataciones

Repetimos el proceso para el año $2013 \Rightarrow i = 4$ con el valor de la tendencia $T_{2013} = T_4 = 26{,}2 - 6 \cdot 4 = 2{,}2$ y podremos estimar:

$y_{2013,1qua} = 2{,}2 + 3{,}4 = 5{,}6$ contractaciones

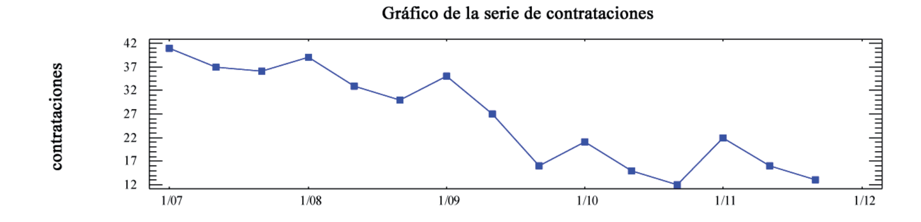
$y_{2013,2qua} = 2{,}2 - 0{,}6 = 1{,}6$ contractaciones

$y_{2013,3qua} = 2{,}2 - 2{,}8 = - 0{,}6$ contractaciones

c) Realiza el análisis de la serie por el método de las medias móviles, estimando también las componentes de la serie.

Gráfico y análisis de la serie

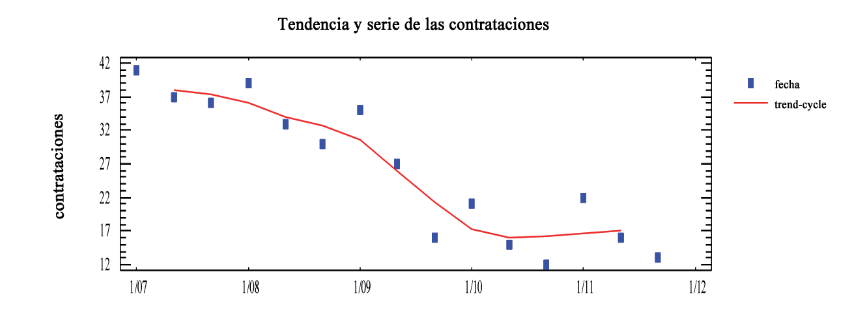
Para analizar la serie por el método de las medias móviles, hay que fijarse en la gráfica de la serie que nos permite ver que tiene una periodicidad anual.



Cálculo de la tendencia

Para calcular la tendencia hay que elegir el número de datos que tomaremos para determinar cada una de las medias. En este caso *p* = 3. A tal fin, escribimos en la tabla siguiente los resultados de las medias móviles que son los valores de la tendencia y la gráfica de su interpretación:

datos	T_k
41	
37	38,00
36	37,33
39	36,00
33	34,00
30	32,67
35	30,67
27	26,00
16	21,33
21	17,33
15	16,00
12	16,33
22	16,67
16	17,00
13	



Se ve un comportamiento a largo plazo decreciente, aunque con los últimos datos parece que empieza a corregirse.

Cálculo de la componente estacional

Para calcular la componente estacional, hay que determinar en primer lugar las medias de los datos originales por períodos. Las denotamos por $\overline{y}_k$.

	2007	2008	2009	2010	2011	$\overline{y}_k$
1.er cuatrimestre	41	39	35	21	22	31,6
2.º cuatrimestre	37	33	27	15	16	25,6
3.er cuatrimestre	36	30	16	12	13	21,4

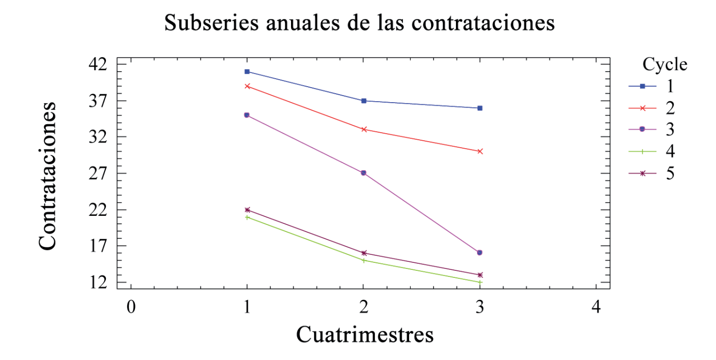
Por otra parte, tomaremos las medias móviles con $p = 3$, ya que es el número de datos por año. En nuestro caso podremos aprovechar los datos ya calculados en el apartado anterior. Los colocaremos distribuidos por años y cuatrimestres.

En la columna de la derecha añadimos las medias de cada cuatrimestre por filas y las denotamos por $\overline{E}_k$.

	2007	2008	2009	2010	2011	$\overline{E}_k$	e_k
1.er cuatrimestre		36,00	30,67	17,33	16,67	25,17	6,43
2.º cuatrimestre	38,00	34,00	26,00	16,00	16,56	26,11	–0,51
3.er cuatrimestre	37,33	32,67	21,33	16,33		26,92	–5,52

Para hallar la componente estacional restamos estas columnas calculadas entre sí. Así, $e_k = \bar{y}_k - \bar{E}_k$ nos permite interpretar el comportamiento de la serie por cuatrimestres.

Es evidente que los primeros cuatrimestres los valores de la serie tienen los mayores valores con 6,43 contrataciones por encima de una hipotética media anual. Destacaremos también el resultado del tercer cuatrimestre con valores inferiores, que se reflejan con el –5,52 contrataciones que veamos en la celda correspondiente. Esta interpretación también es coherente con la gráfica de la serie que hemos visto al comenzar el problema.



Cálculo de la componente errática o residual

Para calcular la componente residual le restaremos a los datos originales la tendencia y la componente estacional, considerando que no lo podremos hacer con el primer y último dato por falta de las medias de la tendencia: $r_{ik} = y_{ik} - T_{ik} - e_k$. En la siguiente tabla se presentan los cálculos y resultados.

	2007	2008	2009	2010	2011
1.er cuat.		39 – 36 – 6,43 = –3,43	35 – 30,67 – 6,43 = –2,1	21 – 17,33 – 6,43 = –2,76	22 – 16,67 – 6,43 = –1,1
2.o cuat.	37 – 38 + 0,51 = –0,49	33 – 34 + 0,51 = –0,49	27 – 26 + 0,51 = 1,51	15 – 16 + 0,51 = –0,49	16 – 16,56 + 0,51 = –0,05
3.er cuat.	36 – 37,33 + 5,52 = 4,19	30 – 32,67 + 5,52 = 2,85	16 – 21,33 + 5,52 = 0,19	12 – 16,33 + 5,52 = 1,19	

Aunque no son cantidades muy elevadas, hemos reseñado en rojo los datos que corresponden al primer cuatrimestre del año 2008 por ser 3,43 menos contrataciones de lo que se espera para la regularidad del fenómeno. También hemos reseñado el dato del tercer cuatrimestre de 2007, que es de 4,19 contrataciones por encima.

d) Compara los resultados del análisis por los dos métodos empleados.

La principal diferencia entre ambos métodos es la posibilidad de hacer previsiones por el método del ajuste, que no podemos hacer con el de las medias móviles, ya

que la ecuación de la recta de tendencia que obtenemos con el método del primer enunciado nos permite obtener valores de tendencia por años cercanos, siempre que consideramos como hipótesis de trabajo que el fenómeno mantendrá un comportamiento similar a los datos reales que tenemos.

Con ambos métodos la tendencia es decreciente.

Si consideramos la componente estacional del método del ajuste

$e_1 = 3{,}4$ $\quad e_2 = -0{,}6$ $\quad e_3 = -2{,}8$

y las comparamos con las del método de las medias móviles

$e_1 = 6{,}43$ $\quad e_2 = -0{,}51$ $\quad e_3 = -5{,}52$

podemos observar que no coinciden en valores absolutos pero sí en la apreciación subjetiva y los valores relativos dan la misma interpretación.

En la componente residual se pueden advertir resultados muy diferentes. También hay que ver que los datos de los últimos dos años son muy inferiores a los anteriores y pueden «romper» un poco el modelo. Ahora bien, como el comportamiento estacional sí era constante, esta componente ha sido más clara.

Bibliografía

Barbancho, A. G., *Estadística Elemental Moderna.* Ed. Ariel Economía, 1982.

Beltrán, J. y Peris, M. J., *Introducció a l'estadística aplicada a les ciències socials*. Servei de Publicacions de la uji. Col·lecció Sapientia, 2013.

Escudero Vallés, R., *Métodos estadísticos aplicados a la economía.* Ed. Ariel Economía, 1994.

Biosca, A.; Espinet, M. J.; Fandos, M. J.; Jimeno, M. y Villagrà, J., *Matemáticas aplicadas a las Ciencias Sociales II*. Barcelona: Edebé, 1999.

Brunet, I., Belzunegui, A. y Pastor, I. *Les tècniques d'investigació social i la seva aplicació.* Universitat Rovira i Virgili, 2000.

Colera, J.; García, R. y Oliveira, M. J. *Matemàtiques aplicades a les Ciències Socials*. Madrid: Anaya, 2003.

Correa, J. C. y González, N., *Gráficos en R*. Universidad Nacional Sede Medellín, 2002.

Fernández Cuesta, C. y Fuentes García, F., *Curso de Estadística Descriptiva. Teoria y práctica*. Ed. Ariel, 1994.

Gracia, F.; Mateu, J. y Vindel, P., *Problemas de Probabilidad y Estadística.*Valencia: Tilde, 1997.

Ibáñez, M. V. y Simó, A., *Apuntes de Estadística para Ciencias Empresariales*. Castellón: uji, 2002.

Kazmier, L., *Estadística aplicada a la administración y a la economía*. Ed. MC Graw-Hill, 3.ª ed., 1998.

Martín, P. y Martín Pliego, J., *Curso Básico de Estadística Económica*. Ed. AC, 3.ª ed., 1991.

Martín Pliego, J., *Introducción a la Estadística Económica y Empresarial*. Ed. AC. Colección Plan Nuevo, 2004.

Meyer, P. L. *Probabilidad y aplicaciones estadísticas,* Ed. Addson-Wesley, 1986.

Monteagudo, M. F. y Paz, J., *Matemáticas aplicadas a las Ciencias Sociales II*. Zaragoza: Luis Vives, 2003.

Montero Lorenzo, J. M., *Estadística para Relaciones Laborales*. Ed. AC, 2003.

Newbold, P. Carlson, W. L. y Thorne, B., *Estadística para administración y economía*. Ed. Prentice Hill, 2007.

Ruíz-Maya Pérez, L. y Martín-Pliego López, F. J., *Fundamentos de Inferencia Estadística,* 3.ª ed., Thomson, 2005.

Sanz, J. A., Bedate, A., Rivas, A. y González, J., *Problemas de Estadística descriptiva empresarial.* Ed. Ariel Economía, 1996.

Spiegel, M., *Estadística.* Ed. Mc. Graw-Hill. Serie Schaum, 1970.

Tomeo Perucha, V. y Uña Juárez, I., *Diez Lecciones de Estadística Descriptiva (Curso Teórico-Práctico).* Ed. AC, 2003.

Triola, M. F., *Estadística Elemental*, Ed. Pearson Educations, 7.ª ed., 2000.

Venables, W. N., Smith , D. M. y The R Development Core Team, *An introduction to R*. isbn 3-900051-12-7, 2008.

Webster, A. L., *Estadística aplicada a los negocios y a la economia*. Ed. MC Graw-Hill, 2000.

Wonnacot, T. H. y Wonnacot, R. J., *Introducción a la Estadística*. Limusa Noriega Editores, 1996.

Zaiats, V., Calle, M. L. y Presas, R., *Probabilitat i Estadística. Exercicis I*. Ed. Eumo, 1998.

Webs

www.ine.es
www. ub.edu/stat/GrupsInnovacio/Statmedia
www.monografias.com